# Sample Complexity and Overparameterization Bounds for Temporal Difference Learning with Neural Network Approximation

Semih Cayci, Siddhartha Satpathi, Niao He, and R. Srikant, *Fellow, IEEE*

***Abstract*— In this paper, we study the dynamics of temporal difference learning with neural network-based value function approximation over a general state space, namely, Neural TD learning. We consider two practically used algorithms, projection-free and max-norm regularized Neural TD learning, and establish the first convergence bounds for these algorithms. An interesting observation from our results is that max-norm regularization can dramatically improve the performance of TD learning algorithms in terms of sample complexity and overparameterization. The results in this work rely on a Lyapunov drift analysis of the network parameters as a stopped and controlled random process.**

***Index Terms*— Reinforcement learning, temporal-difference learning, neural networks, stochastic approximation**

## I. Introduction

Recently, reinforcement learning (RL) algorithms have achieved significant successes in complicated control problems across a broad spectrum of applications including robotics [15], [20], [40], [42], autonomous driving [21], [32], network control [24], [48] and video gaming [26], [33], [34]. An important component of these success stories lies in the power and versatility provided by neural networks in function approximation. Despite the impressive empirical success, the convergence properties of RL algorithms with neural network approximation are not yet fully understood due to their inherent nonconvexity.

In this paper, we investigate the convergence of temporal-difference (TD) learning algorithm equipped with a two-layer fully-connected neural network, namely Neural TD learning, which is an important building block of many RL algorithms. Convergence of TD learning with linear function approximation and least-squares approximation has been established in the literature [6], [43], [49]. On the other hand, it is well-known that using nonlinear approximation may lead to divergence in TD learning [43]. Nonetheless, TD learning with neural network approximation is widely used in practice for policy evaluation because of its simplicity and empirical effectiveness [23], [44]. Therefore, it is important to understand and analyze the convergence properties of Neural TD learning. Recent study of overparameterized networks in the so-called neural tangent kernel (NTK) regime provided important insights in explaining the empirical success of neural networks in supervised learning [2], [3], [12], [13], [16], [17], and thereafter reinforcement learning [8], [10], [35], [47]. Despite the theoretical insights provided by recent studies, there is still a large gap between theory and practice. As we will discuss later, these prior works either consider Neural TD learning in the infinite width limit for finite state spaces, or Neural TD learning with $\ell_2$-projection; neither of which is used in practice. Moreover, explicit characterization of the sample complexity and the amount of overparameterization required for Neural TD learning algorithms to approximate the true value function within arbitrary accuracy has remained elusive, which we address in this paper.

### A. Main Contributions

The paper presents a non-asymptotic analysis of TD learning with two-layer neural network approximation. We elaborate on some of the contributions in this paper below:

- **Analysis of Neural TD learning:** We analyze two practically used Neural TD learning algorithms: (i) vanilla projection-free Neural TD and (ii) max-norm regularized Neural TD. We prove, for the first time, that both algorithms achieve any given target error within a provably rich function class, which is dense in the space of continuous functions over a compact state space. In particular, we establish explicit bounds on the required number of samples, step-size and network width to achieve a given target error.
- **Improved convergence bounds:** We show that projection-free and max-norm regularized Neural TD improve the prior state-of-the-art overparameterization bounds by factors of $1/\epsilon^2$ and $1/\epsilon^6$, respectively, for a given target error $\epsilon$. Notably, we prove that max-norm regularized Neural TD achieves the sharpest overparameterization and sample complexity bounds in the literature, which theoretically supports its empirical effectiveness.

This work was supported in part by the National Science Foundation under Grant CCF 22-07547, CCF 19-34986 and Grant CNS 21- 06801, in part by the Office of Naval Research under Grant N00014-19- 1-2566, and in part by the Swiss National Science Foundation Project Funding under Grant 200021-207343.

S. Cayci and S. Satpathi are with Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (e-mails: {scayci, ssatpth2}@illinois.edu).

N. He is with the Department of Computer Science at ETH Zurich, Zurich 8006, Switzerland (e-mail: niao.he@inf.ethz.ch).

R. Srikant is with c3.ai Digital Transformation Institute, the Department of Electrical and Computer Engineering and Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (e-mail: rsrikant@illinois.edu).

- **Key insights on regularization:** Our analysis reveals that using regularization based on $\ell_\infty$ geometry leads to considerably improved overparameterization and sample complexity bounds compared to the $\ell_2$-regularization over a provably rich function class in the NTK regime.
- **Analytical techniques:** We propose a Lyapunov drift analysis to track the evolution of neural network parameters and the error simultaneously using vector martingale concentration and stopping times. Our analysis can be of independent interest to the analysis of other stochastic approximation and deep RL methods with quadratic loss.

### B. Comparison with Previous Results

Variants of Neural TD learning have been analyzed in the literature. For a quantitative comparison in terms of the required sample complexity and overparameterization bounds to achieve a given target error, please see Table I.

The first result on the convergence of Neural TD learning was presented in [10]. Their work builds upon the analysis in [3], [7], [13], [43] and requires constraining the network parameter within a compact set through the $\ell_2$-projection at each iteration. They prove convergence to a stationary point in a *random* function class $\mathcal{F}_{B,m}$ where $m$ is the network width and $B$ is a given projection radius. Consequently, the algorithm suffers from an approximation error $\epsilon_m = O(\mathbb{E}[\|V - \Pi_{\mathcal{F}_{B,m}} V\|_\mu])$ where $V$ is the value function and $\Pi_{\mathcal{F}_{B,m}}$ denotes the projection onto $\mathcal{F}_{B,m}$. This approximation error is not explicitly bounded in [10], and possibly non-vanishing even with increasing width and projection radius. It is shown in [10], [44] that this variant of Neural TD learning with projection, equipped with a ReLU network of width $O(1/\epsilon^8)$ achieves an error $\epsilon + \epsilon_m$ after $O(1/\epsilon^4)$ iterations. Unlike [10], [44], our Neural TD learning algorithms converge to the *true* value function in a provably rich function class without any approximation error. We show that the algorithms that we consider in this paper achieve improved overparameterization bounds $\widetilde{O}(1/\epsilon^6)$ and $\widetilde{O}(1/\epsilon^2)$ for a given target error $\epsilon$, which improve the existing results by $1/\epsilon^2$ to $1/\epsilon^6$.

In practice, projection-free [26] and max-norm regularized [14], [36], [39] algorithms are often adopted in training neural networks because of their computational efficiency and expressive power, which we consider in this work. In contrast, the Neural TD with $\ell_2$-projection considered in [10], [44] can be computationally expensive for high-dimensional state-spaces as it cannot be performed in parallel.

Projection-free Neural TD learning has also been considered in [1], [8]; however, these works only deal with finite state-space problems in the infinite-width regime, i.e., they do not provide bounds on the amount of overparameterization required. Since these results rely on the positive definiteness of the limiting kernel, the required overparameterization is much larger than the size of the state space which negates the benefits of Neural TD learning over tabular TD learning.

Our work is related to the analysis of (stochastic) gradient descent in the NTK regime. It is shown in [13], [16] that the network parameters trained by gradient descent lie inside a ball around their initialization. However, they require massive overparameterization to ensure the positive definiteness of the neural tangent kernel, which would imply finite state and width much larger than the size of the state space in Neural TD learning. To establish such a result for stochastic gradient descent (and with modest overparameterization) requires additional work, and this problem has been considered for supervised learning tasks in [17], [27]. Specifically, our work builds on the NTK analysis in [17], but deviates from it in the following ways: (a) unlike in classification problems where one has access to the labels, in TD learning, we do not have access to the target function that needs to be approximated, and (b) our loss function is squared-error loss, which has different characteristics than the logistic loss function with exponential tail. These differences make the analysis significantly more complicated.

### C. Notation

For any index set $\mathcal{I}$ and set of vectors $\{b_i \in \mathbb{R}^d : i \in \mathcal{I}\}$, we denote $[b_i]_{i\in\mathcal{I}} \in \mathbb{R}^{d|\mathcal{I}|}$ as the vector that is created by the concatenation of $\{b_i : i \in \mathcal{I}\}$. For an event $\mathcal{E}$ and random variable $X$, we denote $\mathbb{E}[X;\mathcal{E}] = \mathbb{E}[X \cdot \mathbb{I}_\mathcal{E}]$ where $\mathbb{I}_\mathcal{E}$ is the indicator function of the event $\mathcal{E}$. $A^c$ denotes the complement of a set $A$. For any integer $n \geq 1$, $[n] = \{1, 2, \ldots, n\}$. We denote $\ell_2$-norm of a vector $x \in \mathbb{R}^d$ by $\|x\|_2 = \sqrt{\sum_{i\in[d]} |x_i|^2}$ and its $\ell_\infty$-norm by $\|x\|_\infty = \max_{i\in[d]} |x_i|$. For any vector $x \in \mathbb{R}^d$ and $\rho > 0$, $\mathcal{B}_2(x, \rho)$ denotes the ball in $\mathbb{R}^d$ with radius $\rho$ centered at $x$ with respect to the $\ell_2$-norm.

## II. System Model

For simplicity, we consider a Markov reward process $\{(s_t, r_t) : t = 0, 1, \ldots\}$, where the Markov chain $s_t$ takes on values in the state space $\mathcal{S}$, and there is an associated reward $r_t = r(s_t)$ in every time-step for a reward function $r : \mathcal{S} \to [0, 1]$. The process $\{s_t : t \geq 0\}$ evolves according to the transition probabilities $P(s, A) = \mathbb{P}(s_{t+1} \in A | s_t = s)$ for any $s \in \mathcal{S}$, $A \subset \mathcal{S}$ and $t \geq 0$. We assume that the Markov chain $\{s_t : t \geq 0\}$ is an ergodic unichain, therefore there exists a stationary probability distribution $\pi$:

$$\pi(A) = \lim_{t\to\infty} \mathbb{P}(s_t \in A | s_0 = s), \ \forall s \in \mathcal{S}, A \subset \mathcal{S}.$$

The value function associated with the Markov reward process $\{(s_t, r_t) : t \geq 0\}$ is defined as follows:

$$V(s) = \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^t r_t | s_0 = s\Big], \ \forall s \in \mathcal{S}, \tag{1}$$

where $\gamma \in (0, 1)$ is the discount factor. The Bellman operator for this Markov reward process, denoted by $\mathcal{T}$, is defined as follows:

$$\mathcal{T}\widehat{V}(s) = r(s) + \gamma \int_{s'\in\mathcal{S}} \widehat{V}(s') P(s, ds'), \ \forall s \in \mathcal{S}. \tag{2}$$

The value function $V$ is the fixed point of the Bellman operator $\mathcal{T}$: $V(s) = \mathcal{T}V(s)$ for all $s \in \mathcal{S}$. If the state space $\mathcal{S}$ is large, or countably or uncountably infinite, the direct solution of the so-called Bellman equation is computationally inefficient, thus approximation methods are used to evaluate

| Paper | State space | Network width | Sample complexity | Error | Regularization |
|---|---|---|---|---|---|
| Cai et al. [10] | General | $O(\frac{1}{\epsilon^8})$ | $O(\frac{1}{(1-\gamma)^2\epsilon^4})$ | $\epsilon+\epsilon_m$ | $\ell_2$-projection |
| Wang et al. [44] | General | $O(\frac{1}{\epsilon^8})$ | $O(\frac{1}{(1-\gamma)^2\epsilon^4})$ | $\epsilon+\epsilon_\infty$ | $\ell_2$-projection |
| Agazzi & Lu [1] | Finite | $poly(\lvert\mathcal{X}\rvert)$ | $O(\log(1/\epsilon))$ | $\epsilon$ | $poly(\lvert\mathcal{X}\rvert)$ width |
| This paper (PF-NTD) | General | $\widetilde{O}(\frac{1}{\epsilon^6})$ | $O(\frac{1}{(1-\gamma)^4\epsilon^6})$ | $\epsilon$ | Early stopping |
| This paper (MN-NTD) | General | $\widetilde{O}(\frac{1}{\epsilon^2})$ | $O(\frac{1}{(1-\gamma)^2\epsilon^4})$ | $\epsilon$ | Max-norm projection |

TABLE I

THE OVERPARAMETERIZATION AND SAMPLE COMPLEXITY BOUNDS FOR NEURAL TD-LEARNING ALGORITHMS. PF-NTD DENOTES PROJECTION-FREE, MN-NTD DENOTES MAX-NORM REGULARIZED NEURAL TD LEARNING ALGORITHM. $\boldsymbol{\epsilon_m = \mathbb{E}\|V - \Pi_{\mathcal{F}_{B,m}} V\|_\pi}$ DENOTES THE APPROXIMATION ERROR OF THE RANDOM FUNCTION CLASS $\boldsymbol{\mathcal{F}_{B,m}}$ FOR A GIVEN VALUE FUNCTION $\boldsymbol{V}$.

$V$. In this paper, we study the problem of approximating value functions using neural networks given samples from the Markov reward process. We note that the Markov reward process is typically obtained by applying a stationary policy to a controlled Markov process.

For simplicity, we consider independent and identically distributed samples from the stationary distribution $\pi$ of the Markov chain in this paper. Namely, at time $t$, we obtain an observation vector $(s_t, s_t')$ where $s_t \sim \pi$ and $s_t' \sim P(s_t, \cdot)$[1]. We denote $\mathcal{F}_t = \sigma(\{(s_j, s_j') : j = 0, 1, \ldots, t\})$ to be the history up to (including) time $t$. The case where the samples are generated by the Markov reward process can be handled as in [7], [38], [47]. The analysis of MN-NTD under Markovian sampling is provided in Appendix III.

In a broad class of reinforcement learning applications, each state $s \in \mathcal{S}$ is represented by a $d$-dimensional vector $\psi(s)$ where $\psi : \mathcal{S} \to \mathbb{R}^d$. For example, in [26], each state in an Atari game is represented by the corresponding high-dimensional raw image data, while the positions of the players on the board are encoded as a high-dimensional state vector in [33], [34]. For a given trajectory $\{s_t : t \geq 0\}$, we denote the state representations by $x_t = \psi(s_t)$ for $t \geq 0$. We denote the space of state representations as $\mathcal{X} = \{x \in \mathbb{R}^d : x = \psi(s), s \in \mathcal{S}\}$, and use $V(x), r(x), \pi(x)$, etc. to denote the quantities related to a state $\psi^{-1}(x) \in \mathcal{S}$ with a slight abuse of notation. Without loss of generality, we make the following assumption on the state representation, which is commonly used in the neural network literature [3], [10], [17], [27].

*Assumption 1:* For any state $s \in \mathcal{S}$, we assume $\|\psi(s)\|_2 \leq 1$.

In the next subsection, we introduce the neural network architecture that will be used to approximate the value function.

### A. Neural Network Architecture for Value Function Approximation

Throughout the paper, we consider the two-layer ReLU network to approximate the value function $V$:

$$\begin{aligned} Q(x; W, a) &= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \varrho(W_i^\top x) \\ &= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \mathbb{I}\{W_i^\top x \geq 0\} W_i^\top x. \end{aligned} \tag{3}$$

[1]We use the notation $\sim$ to denote that the variable on the left of the symbol is drawn from a distribution to the right of the symbol.

where $\varrho(z) = \max\{0, z\} = z \cdot \mathbb{I}\{z \geq 0\}$ is the ReLU activation function, $a_i \in \mathbb{R}$ and $W_i \in \mathbb{R}^d$ for $i \in [m]$. We include a bias term in $W_i$'s, and express $x$ as $(x, c)$ for a constant $c \in (0, 1)$.

**Symmetric initialization:** The NTK regime is established by random initialization, and various initialization schemes are used, as a common example $a_i \overset{\text{iid}}{\sim} \text{Unif}\{-1, +1\}$ and $W_i(0) \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ [17], [28]. In this paper, we consider an almost-equivalent symmetric variant of this initialization for the sake of simplicity, which was proposed in [4]: $a_i = -a_{i+m/2} \overset{\text{iid}}{\sim} \text{Unif}\{-1, +1\}$ and $W_i(0) = W_{i+m/2}(0) \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ independent and identically distributed over $i = 1, 2, \ldots, m/2$, and independent from each other. The additional benefit of the symmetric initialization is that it provides $Q(x; W(0), a) = 0$ with probability 1 for all $x \in \mathcal{X}$. Without symmetric initialization, $\lim_{m\to\infty} Q(x; W(0), a)$ acts like a random noise term, which leads to an additional approximation error [41]. We fix $a_i$ as initialized, and update $W_i(t)$ by using gradient steps, as in [3], [13], [22]. The sigma field generated by $\{a_i, W_i(0) : i \in [m]\}$ is denoted as $\mathcal{F}_{init}$.

**Function class:** Define the space

$$\mathcal{H} = \left\{ v : \mathbb{R}^d \to \mathbb{R}^d \;\middle|\; \mathbb{E}\left[\|v(w_0)\|_2^2\right] < \infty, w_0 \sim \mathcal{N}(0, I_d) \right\}.$$

We assume that the value function $V$ lies in the following function class.

*Assumption 2:* There exists a vector $v \in \mathcal{H}$ and $\bar{\nu} \geq 0$ such that:

$$V(x) = \mathbb{E}[v^\top(w_0)\phi(x; w_0)], \;\; w_0 \sim \mathcal{N}(0, I_d), \;\; \forall x \in \mathcal{X}, \tag{4}$$

where $\sup_{w\in\mathbb{R}^d} \|v(w)\|_2 \leq \bar{\nu}$ and $\phi(x; w) = \mathbb{I}\{w^\top x \geq 0\}x$.

*Remark 1:* If we replace the condition $\sup_{w\in\mathbb{R}^d} \|v(w)\|_2 \leq \bar{\nu}$ in Assumption 2 by $\mathbb{E}\left[\|v(w_0)\|_2^2\right] < \infty$, then it implies that $V$ belongs to the reproducing kernel Hilbert space (RKHS) induced by the Neural Tangent Kernel (NTK) corresponding to the infinite width neural network given by

$$\begin{aligned} K(x, y) &= \mathbb{E}[\phi(x; w_0)^\top \phi(y; w_0)] \\ &= \mathbb{E}[\mathbb{I}\{w_0^\top x \geq 0\}\mathbb{I}\{w_0^\top y \geq 0\} x^\top y], \end{aligned} \tag{5}$$

with the inner product between functions $f(\cdot) = \mathbb{E}[u^\top(w_0)\phi(\cdot; w_0)]$ and $g(\cdot) = \mathbb{E}[v^\top(w_0)\phi(\cdot; w_0)]$ is defined as $\langle f, g\rangle_{\text{NTK}} = \mathbb{E}[u^\top(w_0)v(w_0)]$ [29]. For a detailed discussion of kernel methods and NTK analysis, see [31], [41], respectively. The above kernel can be shown to be a universal kernel [18] and hence the RKHS induced by the NTK is dense in the space of continuous functions on compact set $\mathcal{X}$

[25]. Therefore, it is possible to replace Assumption 2 by the more general assumption that $V$ is continuous on a compact state space $\mathcal{X}$. In this case, from [18, Theorem 4.3], we know that one can find a function $\tilde{V}$ in the RKHS associated with the NTK, i.e., $\tilde{V}(x) = \mathbb{E}[\tilde{v}^\top(w_0)\phi(x;w_0)]$, $\forall x \in \mathcal{X}$, such that $\sup_w \|\tilde{v}(w)\|_2 \leq \bar{\nu}$ for some finite $\bar{\nu}$ which approximates $V$, where $\bar{\nu}$ depends on the approximation error $\sup_x |V(x) - \tilde{V}(x)|$. If we replace Assumption 2 by the assumption that $V$ is continuous, then the results later can be modified to reflect this approximation error.

*Remark 2:* We note $\sup_{w\in\mathbb{R}^d} \|v(w)\|_2 \leq \bar{\nu}$ in Assumption 2 implies that $\mathbb{E}\big[\|v(w_0)\|_2^2\big] \leq \bar{\nu}$, thus $\bar{\nu}$ is an upper bound on the RKHS norm of $V$ when it lies in the RKHS.

*Remark 3:* It is worth noting the difference between our work and the projection-free TD learning work in [1], [8]. They consider a finite state space in the infinite width limit. For finite $\mathcal{X}$, choosing $m = poly(|\mathcal{X}|)$ guarantees that the kernel $K$ is strictly positive-definite [3], [13], thus in the infinite width limit, the minimum eigenvalue of the limiting kernel is bounded away from zero. By extending the NTK analysis in [13], one can guarantee with further overparameterization that the empirical kernel

$$\widehat{K}_t(x,y) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{W_i(t)^\top x \geq 0\}\mathbb{I}\{W_i(t)^\top y \geq 0\} x^\top y,$$

under TD learning dynamics is also positive definite, thus it can be shown that the network parameters satisfy $W_i(t) \in \mathcal{B}(W_i(0), \rho/\sqrt{m})$ for some $\rho < \infty$ for all $t \geq 1$. However, such a massive overparameterization, i.e., $m = \Omega(|\mathcal{X}|^p)$ for some $p \geq 1$, is not meaningful for TD learning with function approximation, because one may use tabular TD learning directly instead. Thus, we do not seek to make the kernel $K$ positive definite by massive overparameterization in this work since we consider a general (potentially infinite) state space $\mathcal{X}$. Instead, by Assumption 2, we consider functions that can be realized in the RKHS induced by the NTK, and quantify the required overparameterization in terms of $\bar{\nu}$, a bound on the RKHS norm of $V$.

In the next subsection, we present TD learning algorithms to approximate the value function $V$ by a neural network $Q(.;W,a)$.

## III. Neural Temporal Difference Learning Algorithms

For a given function $\mu = [\mu(x)]_{x\in\mathcal{X}}$, we denote the weighted $\ell_2$-norm of any function $\widehat{V}$ as:

$$\|\widehat{V}\|_\mu = \sqrt{\int_{x\in\mathcal{X}} |\widehat{V}(x)|^2 \mu(dx)}.$$

TD learning aims to minimize mean-squared Bellman error, which is defined as follows:

$$\begin{aligned} L(W,a) &= \|Q(W,a) - \mathcal{T}Q(W,a)\|_\pi^2 \\ &= \int_{x\in\mathcal{X}} \Big(Q(x;W,a) - \mathcal{T}Q(x;W,a)\Big)^2 \pi(dx), \end{aligned} \tag{6}$$

for any $W_i \in \mathbb{R}^d, a_i \in \mathbb{R}$ for $i = 1,2,\ldots,m$, where $Q(W,a) = [Q(x;W,a)]_{x\in\mathcal{X}}$, $\pi$ is the stationary distribution of the Markov chain, and $\mathcal{T}$ is the Bellman operator.

Given the initialization $\{(a_i, W_i(0)) : i \in [m]\}$, the parameter update is performed as follows:

$$W(t+1/2) = W(t) + \alpha\Delta_t \nabla_W Q_t(x_t),$$

where $\alpha > 0$ is the step-size, $Q_t(x) = Q(x;W(t),a)$ is the network output at time step $t \geq 0$, and

$$\Delta_t = r_t + \gamma Q_t(x_t') - Q_t(x_t),$$

is the Bellman error. The algorithm is summarized in Algorithm 1. We consider two variants of the Neural TD learning algorithm:

(1) **Projection-free Neural TD learning (PF-NTD):** The network parameters are updated as follows:

$$W(t+1) = W(t+1/2). \tag{7}$$

For regularization, we utilize early stopping, i.e., the number of samples $T$ is chosen as a function of the problem parameters and target error, which we will specify in Theorem 1.

(2) **Max-norm regularized Neural TD learning (MN-NTD):** For a given parameter $R > 0$, let the set of parameters for max-norm regularization be defined as:

$$\mathcal{G}_{m,R}^i = \{W_i \in \mathbb{R}^d : \|W_i - W_i(0)\|_2 \leq \frac{R}{\sqrt{m}}\}, \forall i \in [m]. \tag{8}$$

Then, the network parameters are updated as follows:

$$W_i(t+1) = \Pi_{\mathcal{G}_{m,R}^i} W_i(t+1/2), \forall i \in [m]. \tag{9}$$

where $\Pi_{\mathcal{G}}(\cdot)$ is the projection operator onto set $\mathcal{G}$.

Note that in both PF-NTD and MN-NTD, the algorithm returns $Q(x; \frac{1}{T}\sum_{t<T} W(t))$ as the output. This averaging scheme is common in first-order methods for optimization problems without strong convexity [7], [9], [10].

Max-norm regularization was introduced in [36], [37], and has been widely used in training neural networks [14], [39]. Note that unlike the $\ell_2$-projection in [10], [44], max-norm regularization in (9) can be performed in parallel for all neurons $i \in [m]$, which makes it computationally more feasible. Furthermore, it implies projection onto a well-chosen subset, which leads to much sharper overparameterization and sample complexity bounds for a given value function $V$ [10], [44] as we will show in Theorem 2. Therefore, it is practically used in training neural networks [14], [39]. For TD learning, we will specify the choice of $R$ as a function of the smoothness of the value function $V$ for convergence in Theorem 2.

## IV. Main Results

In the following, we consider a general (possibly infinite) state-space $\mathcal{X}$, and present our main result on the performance of Neural TD learning algorithms described in Section III.

### A. Performance of Projection-Free Neural TD Learning

In the following, we present the sample complexity and overparameterization bounds of PF-NTD. The proof of this result is presented in Section V.

*Theorem 1:* Under Assumptions 1 and 2, for any (possibly infinite) state-space $\mathcal{X}$, target error $\epsilon > 0$ and error probability

**Algorithm 1:** PF/MN-Neural TD Learning

Initialization: $-a_i = a_{i+m/2} \sim \text{Unif}\{-1,+1\}$, $W_i(0) = W_{i+m/2}(0) \sim \mathcal{N}(0, I_d)$, $\forall i \in [\frac{m}{2}]$
**for** $t < T-1$ **do**
  Observe $x_t = \psi(s_t), r_t = r(s_t)$ and $x'_t = \psi(s'_t)$ where $(s_t, s'_t) \overset{\text{iid}}{\sim} \pi \circ P(s_t, \cdot)$
  Compute stochastic semi-gradient: $g_t = \big(r_t + \gamma Q_t(x'_t) - Q_t(x_t)\big)\nabla_W Q_t(x_t)$
  Take a semi-gradient step: $W(t+1/2) = W(t) + \alpha g_t$
  **if** *projection-free* **then** $W(t+1) = W(t+1/2)$;
  **end**
  **if** *max-norm regularization* **then** $W_i(t+1) = \Pi_{\mathcal{G}^i_{m,R}} W_i(t+1/2), \forall i \in [m]$;
  **end**
  Update iterate: $\widehat{W}(t+1) = \Big(1 - \frac{1}{t+2}\Big)\widehat{W}(t) + \frac{1}{t+2}W(t+1)$
**end**
Output: $\overline{Q}_T(x) = Q(x; \widehat{W}(T-1), a)$ for all $x \in \mathcal{X}$

$\delta \in (0,1)$, let $\lambda = \frac{3\bar{\nu}^2}{(1-\gamma)\epsilon\delta}$, $\ell(m,\delta) = 4\sqrt{\log(2m+1)} + \sqrt{\log(1/\delta)}$,

$$m_0 = \frac{16\Big(\bar{\nu} + \big(\lambda + \ell(m_0,\delta)\big)\big(\bar{\nu}+\lambda\big)\Big)^2}{(1-\gamma)^2\epsilon^2},$$

and

$$\alpha_0 = \frac{(1-\gamma)\epsilon^2}{(1+2\lambda)^2}\min\Big\{\frac{\lambda^2}{32\bar{\nu}^2(\sqrt{d}+\sqrt{2\log(m_0/\delta)})^2}, 1\Big\}.$$

Then, for any width $m \geq m_0$, PF-NTD with step-size $\alpha \leq \alpha_0$ yields the following bound after $T = \frac{\bar{\nu}^2}{4\alpha(1-\gamma)\epsilon^2}$ iterations:

$$\mathbb{E}\Big[\big\|\overline{Q}_T - V\big\|_\pi; \mathcal{E}_T\Big] \leq \frac{1}{T}\sum_{t<T}\mathbb{E}[\|Q_t - V\|_\pi; \mathcal{E}_T] + \epsilon \leq 4\epsilon, \tag{10}$$

where $Q_t = [Q_t(x)]_{x\in\mathcal{X}}$, $V = [V(x)]_{x\in\mathcal{X}}$, and the expectation is over the random trajectory and random initialization, and the event $\mathcal{E}_T$ is defined as:

$$\mathcal{E}_T = \Big\{\max_{i\in[m]}\|W_i(t) - W_i(0)\|_2 \leq \frac{\lambda}{\sqrt{m}}, t < T\Big\} \cap E_1,$$

for some $E_1 \in \mathcal{F}_{init}$, which is formally defined in (13) and satisfies $\mathbb{P}(\mathcal{E}_T) > 1 - 4\delta$.

Theorem 1 implies that there exists a set $\mathcal{E}_T$ of trajectories which occurs with probability at least $1-4\delta$ such that Algorithm 1 achieves target error $\epsilon$ under the event $\mathcal{E}_T$ for sufficiently large number of samples and overparameterization. Note that Theorem 1 can be interpreted as $m = \widetilde{O}(T/\delta^2)$ where $T = poly(\bar{\nu}/\delta)O(1/\epsilon^6)$ is the number of samples since the neural network processes one sample per iteration. With this interpretation, we observe that regularization is obtained by overparameterization with respect to $T$, the number of samples, akin to the classical NTK results in the literature [3], [13], [16]. The overparameterization bound has polynomial dependence on the number of samples and does not scale with the size of state-space. Unlike [10], [44], our error bound does not contain any additional approximation error terms.

We have the following remark on the main challenges in the proof of Theorem 1.

*Remark 4:* In [10], projection is applied to the network parameters $W(t)$ in each TD learning iteration to keep $W(t)$ inside a ball of a given radius around the random initialization $W(0)$. In the proof of Theorem 1, we propose methods based on the use of Lyapunov drift coupled with martingale concentration to track the evolution of $\|W_i(t) - W_i(0)\|_2$ and the approximation error $\|Q_t - V\|_\pi$ simultaneously.

### B. Performance of Max-Norm Regularized Neural TD Learning

In the following, we present the overparameterization and sample complexity bounds for MN-NTD.

*Theorem 2:* Under Assumptions 1-2, for any error probability $\delta \in (0,1)$, let $\ell(m,\delta) = 4\sqrt{\log(2m+1)} + \sqrt{\log(1/\delta)}$, and $R > \bar{\nu}$. Then, for any target error $\epsilon > 0$, number of iterations $T \in \mathbb{N}$, network width

$$m > \frac{16\Big(\bar{\nu} + \big(R + \ell(m,\delta)\big)\big(\bar{\nu}+R\big)\Big)^2}{(1-\gamma)^2\epsilon^2},$$

and step-size

$$\alpha = \frac{\epsilon^2(1-\gamma)}{(1+2R)^2},$$

MN-NTD yields the following bound:

$$\mathbb{E}\Big[\|\overline{Q}_T - V\|_\pi; E_1\Big] \leq \frac{(1+2R)\bar{\nu}}{\epsilon(1-\gamma)\sqrt{T}} + 3\epsilon,$$

where $E_1 \in \mathcal{F}_{init}$ holds with probability at least $1-\delta$.

The proof of Theorem 2 is similar to, and simpler than the proof of Theorem 1 because the max-norm projection strictly controls the movement of the parameters. The proof can be found in Appendix II.

### C. Remarks

The above Theorems 1 and 2 provide, to the best of our knowledge, the first explicit characterization of the sample complexity and overparametrization required for PF-NTD and MN-NTD to converge to the true value function with target error $\epsilon$. Below we list some further implications.

*$\ell_2$ vs. $\ell_\infty$ regularizations:* Both PF-NTD and MN-NTD yield improved bounds on $m$ compared to the algorithms in [10], [44] over the provably rich NTK function class (see Table I). A key insight from our analysis is that this improvement is mainly because both PF-NTD and MN-NTD are designed to control $\max_{i\in[m]}\|W_i(t) - W_i(0)\|_2$ via the choice of the stopping time (PF-NTD) or max-norm projection (MN-NTD), while the regularization method in [10], [44] is designed to control $\|W(t) - W(0)\|_2$. Notably, NTD with max-norm regularization achieves the sharpest overparameterization and sample complexity bounds among all NTD variants, which justifies the empirical success of max-norm regularization for training ReLU networks in practice [14], [39].

*Convergence rate:* Regularization of PF-NTD relies on early stopping, whereas MN-NTD utilizes more aggressive max-norm regularization. Without any strict control over $\max_{i\in[m]} \|W_i(t) - W_i(0)\|_2$, PF-NTD requires considerably smaller step-sizes for convergence. Consequently, the sample complexity and required width for PF-NTD to achieve a target error $\epsilon$ are worse than MN-NTD for which larger step-sizes can be chosen.

## V. Analysis of the Neural TD Learning Algorithm

In this section, we will prove Theorem 1. Before starting the proof, let us define a quantity that will be central throughout the proof.

*Definition 1:* For $\lambda$ as given in Theorem 1, let

$$\tau_1 = \inf\Big\{t > 0 : \max_{i\in[m]} \|W_i(t) - W_i(0)\|_2 > \frac{\lambda}{\sqrt{m}}\Big\}, \quad (11)$$

be the stopping time at which there exists $i \in [m]$ such that $W_i(t) \notin \mathcal{B}\big(W_i(0), \lambda/\sqrt{m}\big)$ for the first time.

Since the updates, $g_t$, are random in the Neural TD Learning Algorithm (see Algorithm 1), the stopping time $\tau_1$ is random, which constitutes the main challenge in the proof. As we will show, for any $t < \tau_1$, the drift of $W(t)$ can be controlled. Therefore, we will prove that $\tau_1 > T$ with high probability to prove the error bounds in Theorem 1.

**Proof outline:** Below, we outline the proof steps for Theorem 1.

1) First, we will prove a drift bound for $\|W(t) - \overline{W}\|_2$ which holds for all $t < \tau_1$ where $\overline{W} \in \mathbb{R}^{md}$ is a weight vector such that $\nabla_W^\top Q_0(x)\overline{W} \approx V(x)$ for all $x \in \mathcal{X}$.
2) In the second step, we will use the drift bound obtained in the first step in conjunction with a stopped martingale concentration argument to show that $\tau_1 \geq T$ occurs with high probability, thus the drift bound holds for all $t < T$ under that event.
3) Finally, we will use the drift bound again to show that the approximation error is bounded as in Theorem 1 under the high-probability event considered in Step 2.

### A. Step 1: Lyapunov Drift bound for W(t)

We first prove a drift bound on the weight vector $W(t)$, a common step in the analysis of stochastic gradient descent and TD learning with function approximation [7], [10], [17], [47]. Define the point of attraction as follows:

$$\overline{W} = \Big[W_i(0) + a_i \frac{v\big(W_i(0)\big)}{\sqrt{m}}\Big]_{i\in[m]}, \quad (12)$$

where $W(0)$ is the initial weight vector. Intuitively, by the law of large numbers, $\lim_{m\to\infty} \nabla_W^\top Q_0(x)\overline{W} = V(x)$ for any $x \in \mathcal{X}$ under the symmetric initialization (see (32) for details). For error probability $\delta \in (0,1)$, recall that we define $\ell(\delta, m) = 4\sqrt{\log(m+1)} + \sqrt{\log(1/\delta)}$, and let

$$E_1 = \Big\{ \sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{|W_i^\top(0)x| \leq \frac{\lambda}{\sqrt{m}}\} \leq \frac{\lambda + \ell(m,\delta)}{\sqrt{m}}\Big\}, \quad (13)$$

and $\mathcal{E}_t = E_1 \cap \{t < \tau_1\}$ for any $t < T$.

The following key proposition is used to establish the drift bound.

*Proposition 1:* Recall that $\Delta_t = r_t + \gamma Q_t(x_t') - Q_t(x_t)$ is the Bellman error. Under Assumptions 1-2, we have the following inequalities:

(1) $\mathbb{E}\big[\Delta_t\big(Q_t(x_t) - V(x_t)\big); \mathcal{E}_t\big] \leq -(1-\gamma)z_t^2$.
(2) $\mathbb{E}\big[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t\big] \leq \frac{4\bar{\nu}}{\sqrt{m}} z_t$,
(3) For $\ell(m,\delta)$ defined in Theorem 1:

$$\mathbb{E}\big[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{W}; \mathcal{E}_t\big] \leq \frac{4\big(\bar{\nu} + \lambda\big)\big(\lambda + \ell(m,\delta)\big) z_t}{\sqrt{m}}, \quad (14)$$

where $z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}$ and $\mathbb{E}$ is the expectation over random initialization and trajectory.

The proof of Proposition 1 is given in Appendix I. The first inequality in Proposition 1 follows from the fact that the Bellman operator $\mathcal{T}$ is a contraction with respect to $\|.\|_\pi$, and $V$ is the fixed point of $\mathcal{T}$ [43]. The second inequality holds since $\nabla_W^\top Q_0(x)\overline{W}$ turns into an empirical estimate of $V$ with $m/2$ iid samples, where the variance of each term is at most $\bar{\nu}^2$. The last inequality is the most challenging one as it reflects the evolution of the network output over TD learning steps, and it is essential to have $W_i(t) \in \mathcal{B}(W_i(0), \lambda/\sqrt{m})$ to prove that part.

Now we present the main drift bound for the TD update.

*Lemma 1 (Drift Bound):* For any $t \geq 0$, let $\mathbb{E}_t[.] = \mathbb{E}[.|\mathcal{F}_{t-1}]$ with $\mathcal{F}_{-1} = \mathcal{F}_{init}$. Then, we have the following inequalities:

$$\begin{aligned}\mathbb{E}[\mathbb{E}_t\|W(t+1) - \overline{W}\|_2^2; t < \tau_1] \leq \mathbb{E}[\|W(t) - \overline{W}\|_2^2; t < \tau_1] \\ - 2\alpha(1-\gamma)z_t^2 + \alpha^2(1+2\lambda)^2 \\ + 8\alpha z_t\Big(\frac{\bar{\nu} + \big(\bar{\nu} + \lambda\big)\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}}\Big),\end{aligned} \quad (15)$$

where $\overline{W}$ is as defined in (12), $z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}$. Lemma 1 implies that for $t < \tau_1$, i.e., as long as $W_i(t) \in \mathcal{B}(W_i(0), \lambda/\sqrt{m})$ for all $i \in [m]$, the drift can be made negative by sufficiently large width $m$ and sufficiently small step-size $\alpha$.

*Proof:* Recall that

$$\begin{aligned} g_t &= \big(r_t + \gamma Q_t(x_t') - Q_t(x_t)\big)\nabla_W Q_t(x_t) \\ &= \Delta_t \nabla_W Q_t(x_t),\end{aligned} \quad (16)$$

is the semi-gradient, where $\Delta_t = r_t + \gamma Q_t(x_t') - Q_t(x_t)$ is the Bellman error. Since $W(t+1) = W(t) + \alpha g_t$, we have the following relation:

$$\|W(t+1) - \overline{W}\|_2^2 = \|W(t) - \overline{W}\|_2^2 + 2\alpha\Big[g_t^\top\big(W(t) - \overline{W}\big)\Big] + \alpha^2\|g_t\|_2^2.$$

We can write the expected drift in the following form:

$$\begin{aligned}\mathbb{E}_t[\|W(t+1) - \overline{W}\|_2^2; \mathcal{E}_t] = \|W(t) - \overline{W}\|_2^2 \mathbb{I}_{\mathcal{E}_t} \\ + 2\alpha \underbrace{\mathbb{E}_t[g_t^\top](W(t) - \overline{W})}_{(i)} \mathbb{I}_{\mathcal{E}_t} + \alpha^2 \underbrace{\mathbb{E}_t\|g_t\|_2^2}_{(ii)} \mathbb{I}_{\mathcal{E}_t}.\end{aligned} \quad (17)$$

**Bounding (i) in** (17)**:** In order to bound (i), we expand it as follows. For any $t < \tau_1$:

$$\begin{aligned}\mathbb{E}_t[g_t^\top \big(W(t) - \overline{W}\big)] &= \mathbb{E}_t[\Delta_t \cdot \big(Q_t(x_t) - V(x_t)\big)] \\ &+ \mathbb{E}_t[\Delta_t \cdot \big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big)] \\ &+ \mathbb{E}_t[\Delta_t \cdot \big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{W}], \quad (18)\end{aligned}$$

Then, we obtain the inequality in Lemma 1 by applying Proposition 1.

**Bounding (ii) in** (17)**:** The next argument follows the proof of [10, Lemma 4.5]:

$$\begin{aligned}\|g_t\|_2 &= \|(r_t + \gamma Q_t(x_t') - Q_t(x_t))\nabla_W Q_t(x_t)\|_2, \\ &\leq |r_t + \gamma Q_t(x_t') - Q_t(x_t)|, \\ &\leq 1 + 2\max_{x\in\mathcal{X}} |Q_t(x)|, \\ &\leq 1 + 2\|W(t) - W(0)\|_2 \leq 1 + 2\lambda,\end{aligned} \quad (19)$$

where the first inequality follows since $\|\nabla_W Q_t(x)\|_2 \leq 1$ for any $t, x$, the second inequality follows since $r(x) \in [0,1]$ for all $x \in \mathcal{X}$, and the last inequality holds since $|Q_t(x)| = |Q_t(x) - Q_0(x)| \leq \|W(t) - W(0)\|_2$ and $t < \tau_1$. Consequently, $\|g_t\|_2^2 \leq (1+2\lambda)^2$.

The result in (15) immediately follows by combining these two bounds. ∎

### B. Step 2: Stopping time $\tau_1 \geq \mathsf{T}$ with high probability

Now, we will use the drift result in Step 1 to show that $\tau_1 \geq T$ with high probability.

*Lemma 2:* Under Assumptions 1-2, we have:

$$\tau_1 = \inf\Big\{t > 0 : \max_{i\in[m]} \|W_i(t) - W_i(0)\|_2 > \frac{\lambda}{\sqrt{m}}\Big\} \geq T,$$

with probability at least $1 - \delta$.

*Proof:* First, invoking Lemma 1 with the values for $T$, $\lambda$ and $m$ specified in Theorem 1, we have the following inequality for any $t$:

$$\begin{aligned}\mathbb{E}[\mathbb{E}_t\|W(t+1) - \overline{W}\|_2^2; \mathcal{E}_t] &\leq \mathbb{E}[\|W(t) - \overline{W}\|_2^2; \mathcal{E}_t] \\ &- 2\alpha(1-\gamma)z_t^2 + 2\alpha(1-\gamma)\epsilon^2 + 4\alpha(1-\gamma)\epsilon z_t, \quad (20)\end{aligned}$$

where $z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}$. The step-size $\alpha$ is chosen sufficiently small so that, by (19), $\alpha^2\|g_t\|^2 \leq 2\alpha(1-\gamma)\epsilon^2$.

*Claim 1:* Telescoping sum of (20) over $t < T$ yields:

$$0 \leq \bar{\nu}^2 - 2\alpha(1-\gamma)\sum_{t<T}(z_t - \epsilon)^2 + 4\alpha(1-\gamma)\epsilon^2 T.$$

*Proof:* [Proof of Claim 1] Recall the notation $\mathbb{E}_t[.] = \mathbb{E}[.|\mathcal{F}_{t-1}]$. Let $\overline{\delta}(t) = W(t) - \overline{W}$ and $\zeta_T$ be defined as:

$$\begin{aligned}\zeta_T &= \sum_{t<T}\Big(\mathbb{E}_t[\|\overline{\delta}(t+1)\|_2^2]\mathbb{I}_{\mathcal{E}_t} - \|\overline{\delta}(t+1)\|_2^2\mathbb{I}_{\mathcal{E}_{t+1}}\Big), \\ &\geq \sum_{t<T}\mathbb{I}_{\mathcal{E}_t}\Big(\mathbb{E}_t[\|\overline{\delta}(t+1)\|_2^2] - \|\overline{\delta}(t+1)\|_2^2\Big) = \zeta_T',\end{aligned}$$

for $T \geq 1$ with $\zeta_0 = \zeta_0' = 0$, where the inequality holds since $\mathbb{I}_{\mathcal{E}_{t+1}} \leq \mathbb{I}_{\mathcal{E}_t}$. Note that $\zeta_T'$ is a martingale over the filtration $\{\mathcal{F}_t\}$ since each summand constitutes a martingale difference sequence, and $\mathbb{I}_{\mathcal{E}_t} \in \mathcal{F}_{t-1}$ is predictable and nonnegative. Then, we have:

$$\begin{aligned}\sum_{t<T}\Big(\mathbb{E}_t[\|\overline{\delta}(t+1)\|_2^2] - \|\overline{\delta}(t)\|_2^2\Big)\mathbb{I}_{\mathcal{E}_t} &\geq \zeta_T - \bar{\nu}^2 \\ &+ \underbrace{\|W(T) - \overline{W}\|_2^2\mathbb{I}_{\mathcal{E}_T^c}}_{\geq 0},\end{aligned}$$

which follows from $\|W(0) - \overline{W}\|_2 \leq \bar{\nu}$. Since $\zeta_T \geq \zeta_T'$ for any $T \geq 1$, and $\zeta_T'$ is a martingale with $\zeta_0' = 0$, we have $\mathbb{E}[\zeta_T] \geq \mathbb{E}[\zeta_T'] = 0$. Hence,

$$\sum_{t<T}\Big(\mathbb{E}[\|W(t+1) - \overline{W}\|_2^2; \mathcal{E}_t] - \mathbb{E}[\|W(t) - \overline{W}\|_2^2; \mathcal{E}_t]\Big) \geq -\bar{\nu}^2,$$

and therefore the claim follows. ∎

Applying Claim 1 and Jensen's inequality, we have:

$$\sum_{t<T} z_t \leq 3\epsilon T = \frac{3\bar{\nu}^2}{4\alpha(1-\gamma)\epsilon}. \quad (21)$$

This bound on the total error will be the fundamental quantity in the proof. Now, by using (21), we will show that the event $\mathcal{E}_T' = \{\tau_1 < T\} \cap E_1$ occurs with low probability. For any $i \in [m]$, let $\overline{g}_i(t+1) = W_i(t+1) - W_i(t)$. Then, we have:

$$\begin{aligned}\|W_i(\tau_1) - W_i(0)\|_2 &= \|\sum_{t<\tau_1}\overline{g}_i(t+1)\|_2, \\ &\leq \|\sum_{t<\tau_1}\overline{g}_i(t+1) - \sum_{t<\tau_1}\mathbb{E}_t[\overline{g}_i(t+1)]\|_2 \quad (22) \\ &+ \|\sum_{t<\tau_1}\mathbb{E}_t[\overline{g}_i(t+1)]\|_2. \quad (23)\end{aligned}$$

**Bounding** (22)**:** For any $t$, let

$$D_{i,t} = W_i(t+1) - W_i(t) - \mathbb{E}_t[W_i(t+1) - W_i(t)], \quad (24)$$

which forms a martingale difference sequence with respect to the filtration $\mathcal{F}_t$ since $\mathbb{E}_t[D_{i,t}] = 0$. Let $X_{i,t'} = \sum_{t<t'} D_{i,t}$. Since $D_{i,t}$ is a martingale difference sequence, $X_{i,t}$ is a martingale. Thus, bounding (22) is equivalent to bounding $\|X_{i,\tau_1}\|_2$, under the event $\mathcal{E}_T'$. In order to achieve this, we use a concentration inequality for vector-valued martingales [19, Theorem 2.1], which is given in the following.

*Proposition 2 (Concentration for Vector Martingales):* Consider a martingale difference sequence $\{D_t \in \mathbb{R}^d : t \geq 0\}$, and let $X_T = \sum_{t<T} D_t$. If $\|D_t\|_2 \leq \varsigma$ almost surely for all $t$, then for any $T$ and $\beta > 0$, we have the following inequality:

$$\mathbb{P}\Big(\|X_T\| \geq \sqrt{2}\varsigma\big(\sqrt{d} + \beta\big)\sqrt{T}\Big) \leq \exp(-\beta^2/2). \quad (25)$$

Since $\sup_{x\in\mathcal{X}}|Q_t(x)| \leq \|W(t) - W(0)\|_2 \leq \lambda$ for all $t < \tau_1$, we have $\|D_{i,t}\|_2 \leq \frac{2\alpha(1+2\lambda)}{\sqrt{m}}$. Define the stopped martingale $\widetilde{X}_{i,t} = X_{i,\min\{t,\tau_1\}}$, which is again a martingale with a corresponding martingale difference sequence $\widetilde{D}_{i,t}$ that satisfies $\|\widetilde{D}_{i,t}\|_2 \leq \|D_{i,t}\|_2$ [45]. Since

$$\|X_{i,\tau_1}\|_2 \cdot \mathbb{I}_{\mathcal{E}_T'} \leq \|\widetilde{X}_{i,T}\|_2,$$

the following inequality holds:

$$\mathbb{P}\Big(\|X_{i,\tau_1}\|_2 \geq \sqrt{2}(\sqrt{d}+\beta)\frac{2\alpha(1+2\lambda)\sqrt{T}}{\sqrt{m}}; \mathcal{E}_T'\Big) \leq e^{-\beta^2/2},$$

which follows from

$$\{\|X_{i,\tau_1}\|_2 \geq \sqrt{2}(\sqrt{d}+\beta)\frac{2\alpha(1+2\lambda)\sqrt{T}}{\sqrt{m}}\} \cap \mathcal{E}_T' \subset \{\|\widetilde{X}_{i,T}\|_2 \geq \sqrt{2}(\sqrt{d}+\beta)\frac{2\alpha(1+2\lambda)\sqrt{T}}{\sqrt{m}}\},$$

and

$$\mathbb{P}\Big(\|\widetilde{X}_{i,T}\|_2 \geq \sqrt{2}(\sqrt{d}+\beta)\frac{2\alpha(1+2\lambda)\sqrt{T}}{\sqrt{m}}\Big) \leq e^{-\beta^2/2},$$

by Proposition 2. Therefore, by using union bound:

$$\mathbb{P}(\max_{i\in[m]}\|X_{i,\tau_1}\|_2 > 4\big(\sqrt{d}+\sqrt{\log(\frac{m}{\delta})}\big)\sqrt{\frac{T}{m}}\alpha(1+2\lambda); \mathcal{E}_T') \leq \delta, \quad (26)$$

The step-size $\alpha$ is chosen to satisfy

$$\big(\sqrt{2d}+2\sqrt{\log(m/\delta)}\big)\sqrt{T}\cdot 2\alpha(1+2\lambda) \leq \lambda/2.$$

**Bounding** (23)**:** Note that we can bound (23) as follows:

$$\|\sum_{t<\tau_1}\mathbb{E}_t[\overline{g}_i(t+1)]\mathbb{I}_{\mathcal{E}_T'}\|_2 \leq \sum_{t<\tau_1}\frac{2\alpha\mathbb{I}_{\mathcal{E}_T'}}{\sqrt{m}}\|Q_t - V\|_\pi, \quad (27)$$

for all $i \in [m]$ under $\mathcal{E}_T'$ since $\sup_{i,t,x}\|\nabla_{W_i}Q_t(x)\|_2 \leq 1/\sqrt{m}$ (see Claim 3 and Remark 5 for details). The expectation of the RHS above is bounded as follows:

$$\frac{2\alpha}{\sqrt{m}}\mathbb{E}[\sum_{t<\tau_1}\|Q_t - V\|_\pi\mathbb{I}_{\mathcal{E}_T'}] \leq \frac{2\alpha}{\sqrt{m}}\sum_{t<T}\mathbb{E}[\|Q_t - V\|_\pi; \mathcal{E}_t] \leq \frac{2\alpha}{\sqrt{m}}\sum_{t<T}z_t,$$

by the law of iterated expectations as the event $\{t<\tau_1\}\cap E_1 \in \mathcal{F}_{t-1}$ as $\|W_i(t) - W_i(0)\| \in \mathcal{F}_{t-1}$. Note that the RHS of the previous inequality is upper bounded by (21). Therefore, we have:

$$\frac{2\alpha}{\sqrt{m}}\mathbb{E}[\sum_{t<\tau_1}\|Q_t - V\|_\pi; \mathcal{E}_T'] \leq \frac{6T\epsilon\alpha}{\sqrt{m}}.$$

Hence, we have the following:

$$\bigcup_{i\in[m]}\Big\{\|\sum_{t<\tau_1}\mathbb{E}_t[\overline{g}_i(t+1)]\mathbb{I}_{\mathcal{E}_T'}\|_2 > \frac{6\alpha T\epsilon}{\sqrt{m\delta}}\Big\}\cap\mathcal{E}_T' \subset \Big\{\sum_{t<\tau_1}\frac{2\alpha\|Q_t - V\|_\pi\mathbb{I}_{\mathcal{E}_T'}}{\sqrt{m}} > \frac{6\alpha T\epsilon}{\sqrt{m\delta}}\Big\},$$

which implies that

$$\mathbb{P}(\bigcup_{i\in[m]}\Big\{\|\sum_{t<\tau_1}\mathbb{E}_t[\overline{g}_i(t+1)]\|_2 > \frac{6\alpha T\epsilon}{\sqrt{m\delta}}\Big\}; \mathcal{E}_T') \leq \delta, \quad (28)$$

by Markov's inequality. Now, using (26) and (28) in (22) and (23), we conclude that $\mathbb{P}(\mathcal{E}_T') \leq 2\delta$. Since $\mathcal{E}_T^c = \mathcal{E}_T' \cup E_1^c$ and $\mathbb{P}(E_1^c) \leq \delta$ by Claim 2, we conclude that $\mathcal{E}_T$ holds with probability at least $1-3\delta$.

■

### C. Step 3: Error bound

In Step 2, we have shown that the event $\{\tau_1 \geq T\}$ occurs with high probability. Since $\mathcal{E}_T = \{\tau_1 \geq T\} \cap E_1 \subset \mathcal{E}_t$ for any $t < T$, we have the following inequality:

$$\mathbb{E}[\|Q_t - V\|_\pi; \mathcal{E}_T] \leq \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_T]} \leq z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]},$$

for any $t < T$. Consequently, by using (21) and Jensen's inequality, we have:

$$\mathbb{E}[\|\frac{1}{T}\sum_{t<T}Q_t - V\|_\pi; \mathcal{E}_T] \leq \frac{1}{T}\sum_{t<T}\mathbb{E}[\|Q_t - V\|_\pi; \mathcal{E}_T] \leq \frac{1}{T}\sum_{t<T}z_t \leq 3\epsilon.$$

In the final step, by following similar steps as [10], we use Proposition 3 in Appendix I to show the proximity of $\overline{Q}_T$ and $\frac{1}{T}\sum_{t<T}Q_t$ to $\nabla_W^\top Q_0\overline{W}$, and conclude that $\mathbb{E}[\|\overline{Q}_T - \frac{1}{T}\sum_{t<T}Q_t\|_\pi; \mathcal{E}_T] \leq \epsilon$, which implies $\mathbb{E}[\|\overline{Q}_T - V\|_\pi; \mathcal{E}_T] \leq 4\epsilon$ by triangle inequality.

## VI. Conclusion

In this paper, we analyzed two practically used TD learning algorithms with neural network approximation, and established non-asymptotic bounds on the required number of samples and network width to achieve any given target error within a provably rich function class. By using a novel Lyapunov drift analysis of stopped and controlled random processes, we have shown for the first time that projection-free Neural TD learning can achieve arbitrarily small target error. In addition, we proved that max-norm regularized Neural TD learning achieves the state-of-the-art complexity bounds, which theoretically supports its empirical effectiveness in ReLU networks. One key insight from our analysis is that $\ell_\infty$-regularization yields improved results in the NTK regime compared to $\ell_2$-regularization. The extension of this work to other reinforcement learning algorithms, such as Q-learning and policy gradient methods, and different neural network architectures, such as multi-layer and convolutional networks, is left for future investigation. The function class that we considered in this paper is realizable by the neural tangent kernel. Recently, it was shown that neural networks trained by using gradient descent methods are able to learn functions that cannot be learned by kernel methods [46]. As such, an important open problem is to analyze the performance of Neural TD learning beyond the NTK regime.

## References


[1] Andrea Agazzi and Jianfeng Lu. Temporal-difference learning for nonlinear value function approximation in the lazy training regime. *arXiv preprint arXiv:1905.10917*, 2019.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, volume 32, pages 6158–6169, 2019.

[3] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[4] Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*, 2019.

[5] Dimitri P Bertsekas. Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*, 2011.

[6] Dimitri P Bertsekas. Temporal difference methods for general projected equations. *IEEE Transactions on Automatic Control*, 56(9):2128–2139, 2011.

[7] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692. PMLR, 2018.

[8] David Brandfonbrener and Joan Bruna. Geometric insights into the convergence of non-linear td learning. In *International Conference on Learning Representations*, 2020.

[9] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[10] Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[11] Semih Cayci, Niao He, and R Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *arXiv preprint arXiv:2206.00833*, 2022.

[12] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2937–2947, 2019.

[13] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

[14] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR, 2013.

[15] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.

[16] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[17] Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.

[18] Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*, 2019.

[19] Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

[20] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pages 651–673. PMLR, 2018.

[21] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[22] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[23] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[24] Nguyen Cong Luong, Dinh Thai Hoang, Shimin Gong, Dusit Niyato, Ping Wang, Ying-Chang Liang, and Dong In Kim. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials*, 21(4):3133–3174, 2019.

[25] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

[26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[27] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

[28] Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

[29] Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561. IEEE, 2008.

[30] Siddhartha Satpathi, Harsh Gupta, Shiyu Liang, and R Srikant. The role of regularization in overparameterized neural networks. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4683–4688. IEEE, 2020.

[31] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2018.

[32] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[33] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[34] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[35] Justin Sirignano and Konstantinos Spiliopoulos. Asymptotics of reinforcement learning with neural networks. *arXiv preprint arXiv:1911.07304*, 2019.

[36] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2005.

[37] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

[38] R. Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

[39] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[40] Lei Tai, Giuseppe Paolo, and Ming Liu. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 31–36. IEEE, 2017.

[41] Matus Telgarsky. Deep learning theory lecture notes. `https://mjt.cs.illinois.edu/dlt/`, 2021. Version: 2021-02-14 v0.0-1dabbd4b (pre-alpha).

[42] Stephen Tian, Frederik Ebert, Dinesh Jayaraman, Mayur Mudigonda, Chelsea Finn, Roberto Calandra, and Sergey Levine. Manipulation by feel: Touch-based control with deep predictive models. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 818–824. IEEE, 2019.

[43] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[44] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

[45] David Williams. *Probability with martingales*. Cambridge university press, 1991.

[46] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[47] Pan Xu and Quanquan Gu. A finite-time analysis of Q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR, 2020.

[48] Zhiyuan Xu, Jian Tang, Jingsong Meng, Weiyi Zhang, Yanzhi Wang, Chi Harold Liu, and Dejun Yang. Experience-driven networking: A deep reinforcement learning based approach. In *IEEE INFOCOM 2018-IEEE conference on computer communications*, pages 1871–1879. IEEE, 2018.

[49] Huizhen Yu and Dimitri P Bertsekas. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7):1515–1531, 2009.

## APPENDIX I
## PROOFS OF SECTION V

Throughout the section, we will use the following results extensively.

*Claim 2 (Lemma 4.1 in [30]):* For any $\delta \in (0,1)$ and $m \in \mathbb{N}$, let

$$\ell_0(m,\delta) = \sqrt{8d\log(m+1)} + \sqrt{8\log(1/\delta)}.$$

Then, for any $\epsilon > 0$, $m > 10$, if $W_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ for all $i \in [m]$, we have:

$$\sup_{x:\|x\|_2\leq 1} \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{|W_i^\top x| \leq \epsilon\} \leq \sqrt{\frac{2}{\pi}}\epsilon + \frac{\ell_0(m,\delta)}{\sqrt{m}}, \quad (29)$$

with probability at least $1-\delta$ over the random initialization.

*Claim 3:* For any $W \in \mathbb{R}^{md}$ and $a_i \in \{-1,1\}$ for $i \in [m]$, we have:

$$\|\nabla_W Q(x;W,a)\|_2 \leq 1,$$
$$\|\nabla_{W_i} Q(x;W,a)\|_2 \leq 1/\sqrt{m},\ \forall i \in [m],$$

for any $x \in \mathcal{X}$.

*Proof:* Note that for any $i \in [m]$ and $x \in \mathcal{X}$,

$$\nabla_{W_i} Q(x;W,a) = \frac{1}{\sqrt{m}} a_i \mathbb{I}\{W_i^\top x \geq 0\} x,$$

which implies $\|\nabla_{W_i} Q(x;W,a)\|_2 \leq \frac{\|x\|_2}{\sqrt{m}} \leq \frac{1}{\sqrt{m}}$ since $a_i \in \{-1,+1\}$ for all $i \in [m]$ which is automatically satisfied by the symmetric initialization, and $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}$. Similarly,

$$\nabla_W Q(x;W,a) = \Big[\frac{1}{\sqrt{m}} a_i \mathbb{I}\{W_i^\top x \geq 0\} x\Big]_{i\in[m]},$$

which directly implies $\|\nabla_W Q(x;W,a)\|_2^2 \leq \|x\|_2^2 \leq 1$ ■

*Claim 4 (Lemma 6.3.1 in [5]):* For any $\widehat{V} = [\widehat{V}(s)]_{s\in\mathcal{S}}$,

$$\|\mathcal{T}\widehat{V}\|_\pi \leq \gamma \cdot \|\widehat{V}\|_\pi,$$

where $\mathcal{T}$ is the Bellman operator.

*Claim 5:* For any $t \geq 0$, we have:

$$\sqrt{\mathbb{E}_t[(\mathcal{T}Q_t(x_t) - Q_t(x_t))^2]} \leq (1+\gamma)\|Q_t - V\|_\pi.$$

*Proof:* For any $x \in \mathcal{X}$, we have $\mathcal{T}V(x) = V(x)$ since $V$ is the fixed point of the Bellman operator $\mathcal{T}$. Therefore, we have:

$$\sqrt{\mathbb{E}_t[(\mathcal{T}Q_t(x_t) - Q_t(x_t))^2]} = \sqrt{\mathbb{E}_t[(\mathcal{T}Q_t(x_t) - \mathcal{T}V(x_t) - Q_t(x_t) + V(x_t))^2]}.$$

Since $V(x), Q_t(x) \in \mathcal{F}_{t-1}$ for any given $x \in \mathcal{X}$, the expectation $\mathbb{E}_t$ is over $(x_t, x_t')$, thus we have:

$$\sqrt{\mathbb{E}_t[(\mathcal{T}Q_t(x_t) - \mathcal{T}V(x_t) - Q_t(x_t) + V(x_t))^2]} = \|\mathcal{T}Q_t - \mathcal{T}V - Q_t + V\|_\pi.$$

By triangle inequality, the above inequality implies the following:

$$\sqrt{\mathbb{E}_t[(\mathcal{T}Q_t(x_t) - Q_t(x_t))^2]} \leq \|\mathcal{T}Q_t - \mathcal{T}V\|_\pi + \|Q_t - V\|_\pi.$$

Since $\mathcal{T}$ is a contraction over $\|.\|_\pi$ by Claim 4 with modulus $\gamma \in (0,1)$,

$$\|\mathcal{T}Q_t - \mathcal{T}V\|_\pi \leq \gamma\|Q_t - V\|_\pi,$$

which implies the result. ■

*Remark 5:* Note that for any sequence of predictable $\mathbb{R}^k$-valued ($k \geq 1$) functions $h_t \in \mathcal{F}_{t-1}$ which does not depend on $x_t'$, we have the following identity:

$$\mathbb{E}_t[\Delta_t h_t(x_t)] = \int_{\mathcal{X}} \Big(\int_{\mathcal{X}} \Delta_t \pi(dx_t')\Big) h(x_t)\pi(dx_t), = \mathbb{E}_t[\big(\mathcal{T}Q_t(x_t) - Q_t(x_t)\big)h_t(x_t)].$$

We use this identity extensively in the analysis throughout this work, mainly in conjunction with Claim 5. Some examples for this are as follows: $h_t(x) = \nabla_W Q_t(x)$ which leads to $g_t = \mathbb{E}_t[\Delta_t h_t(x_t)]$ and $h_t(x) = \nabla_{W_i} Q_t(x)$ which leads to $W_i(t+1) - W_i(t) = \alpha\mathbb{E}_t[\Delta_t h_t(x_t)]$, where other instances such as $h_t(x) = Q_t(x) - V(x)$ show up in the following analysis as well.

### A. Proof of Proposition 1

**Part (1)** Let $\ell(\delta,m) = 2\sqrt{\log(2m+1)} + \sqrt{\log(1/\delta)}/2$,

$$E_1 = \Big\{\sup_{x\in\mathcal{X}} \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{|W_i^\top(0)x| \leq \epsilon\} \leq \sqrt{\frac{2}{\pi}}\epsilon + \frac{\ell(m,\delta)}{\sqrt{m}}\Big\},$$

For any $t < T$, let $\mathcal{E}_t = E_1 \cap \{t < \tau_1\}$. (1) For any $t < \tau_1$, we have the following inequality:

$$\mathbb{E}_t[\Delta_t\big(Q_t(x_t) - V(x_t)\big)] \leq -(1-\gamma)\|Q_t - V\|_\pi^2.$$

*Proof:* The proof follows the strategy first proposed in [43], and then used for the convergence proofs in [7], [10], [47]. Let $\mathbb{E}_t[.] = \mathbb{E}[.|\mathcal{F}_{t-1}]$, i.e., the expectation is over $(x_t, x_t')$. Then, we have

$$\mathbb{E}_t[\Delta_t(Q_t(x_t) - V(x_t))] = \mathbb{E}_t[(\mathcal{T}Q_t(x_t) - Q_t(x_t))(Q_t(x_t) - V(x_t))],$$

by taking expectation over $x_t'$ first, which implies the following:

$$\mathbb{E}_t\Big[\big(\mathcal{T}Q_t(x_t) - Q_t(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)\Big] = \mathbb{E}_t\Big[\big(\mathcal{T}Q_t(x_t) - \mathcal{T}V(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)\Big] - \mathbb{E}_t\Big[\big(Q_t(x_t) - V(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)\Big],$$

since $\mathcal{T}V(x) = V(x)$ for any $x \in \mathcal{X}$. Therefore, we have:

$$\mathbb{E}_t\Big[\big(\mathcal{T}Q_t(x_t) - Q_t(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)\Big] \leq \eta_t - \|Q_t - V\|_\pi^2.$$

where $\eta_t = \mathbb{E}_t[\big(\mathcal{T}Q_t(x_t) - \mathcal{T}V(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)]$. Since $\|.\|_\pi$ defines a norm, by Cauchy-Schwarz inequality, we have:

$$\begin{aligned}\eta_t &= \mathbb{E}_t[\big(\mathcal{T}Q_t(x_t) - \mathcal{T}V(x_t)\big)\big(Q_t(x_t) - V(x_t)\big)] \\ &\leq \|\mathcal{T}Q_t - \mathcal{T}V\|_\pi \cdot \|Q_t - V\|_\pi.\end{aligned}$$

From Claim 4, we have $\|\mathcal{T}Q_t - \mathcal{T}V\|_\pi \leq \gamma\|Q_t - V\|_\pi$, which implies the result. ■

**Part (2)** For any $t$, we have:

$$\mathbb{E}[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t] \leq \frac{4\overline{\nu}}{\sqrt{m}}\sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}.$$

*Proof:* Let $\nabla_W^\top Q_0 \overline{W} = \big[\nabla_W^\top Q_0(x)\overline{W}\big]_{x\in\mathcal{X}}$. Then, for any $t$, we have:

$$\begin{aligned}&\mathbb{E}_t[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big)] \\ &\quad = \mathbb{E}_t[\big(\mathcal{T}Q_t(x_t) - Q_t(x_t)\big)\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big)]\end{aligned}$$

By using Cauchy-Schwarz inequality, we have:

$$\begin{aligned}&\mathbb{E}_t[\big(\mathcal{T}Q_t(x_t) - Q_t(x_t)\big)\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big)] \\ &\quad \leq \|\mathcal{T}Q_t - Q_t\|_\pi\sqrt{\mathbb{E}_t[(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W})^2]}.\end{aligned}$$

Then, by using Claim 5,

$$\|\mathcal{T}Q_t - Q_t\|_\pi \leq (1+\gamma)\|Q_t - V\|_\pi.$$

By the law of iterated expectations,

$$\begin{aligned}&\mathbb{E}[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t] \\ &\quad \leq (1+\gamma)\mathbb{E}[\|Q_t - V\|_\pi \mathbb{I}_{\mathcal{E}_t}\|V - \nabla_W^\top Q_0\overline{W}\|_\pi],\end{aligned}$$

since $\mathbb{I}_{\mathcal{E}_t} \in \mathcal{F}_{t-1}$. Hence, by Cauchy-Schwarz inequality, we have the following:

$$\begin{aligned}&\mathbb{E}[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t] \\ &\quad \leq 2\sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}\sqrt{\mathbb{E}[\|V - \nabla_W^\top Q_0\overline{W}\|_\pi^2]}. \quad (30)\end{aligned}$$

In the following, we will bound $\sqrt{\mathbb{E}[\|V - \nabla_W^\top Q_0\overline{W}\|_\pi^2]}$. For any $x \in \mathcal{X}$, we have:

$$V(x) - \nabla_W^\top Q_0(x)\overline{W} = \frac{1}{m}\sum_{i=1}^m \Big(V(x) - \widehat{V}_i(x)\Big). \quad (31)$$

where $\widehat{V}_i(x) = \mathbb{I}\{W_i^\top(0)x \geq 0\}v^\top(W_i(0))x$. Recall that $V(x) = \mathbb{E}[\mathbb{I}\{W_i^\top(0)x \geq 0\}v^\top(W_i(0))x] = \mathbb{E}[\widehat{V}_i(x)]$ by Assumption 2, which implies

$$\nabla_W^\top Q_0(x)\overline{W} = \frac{1}{m}\sum_{i=1}^m \widehat{V}_i(x) \to V(x) \quad (32)$$

as $m \to \infty$ almost surely for any $x \in \mathcal{X}$ by the strong law of large numbers since $V_i(x)$ is a bounded random variable for all $i \in [m]$ and $x \in \mathcal{X}$. For any $i \in [m]$, we have:

$$\mathbb{E}[V(x) - \widehat{V}_i(x)] = 0,$$

and for $i, j \in [m/2]$, we have:

$$Cov\Big(\widehat{V}_i(x), \widehat{V}_j(x)\Big) \leq \mathbb{E}[\|v(W_1(0))\|_2^2],$$

if $i = j$, and $Cov\Big(\widehat{V}_i(x), \widehat{V}_j(x)\Big)$ if $i \neq j$. Under symmetric initialization, $W_i(0) = W_{i+m/2}(0)$ for all $i \in [m/2]$. Therefore, by using the above result along with Fubini's theorem, we have:

$$\begin{aligned}&\mathbb{E}\|V - \nabla_W^\top Q_0\overline{W}\|_\pi^2 \\ &\quad = \mathbb{E}\Big[\int_{x\in\mathcal{X}}\Big(\frac{1}{m}\sum_{i=1}^m\Big(V(x) - \widehat{V}_i(x)\Big)\Big)^2\pi(dx)\Big], \\ &\quad \leq \int_{x\in\mathcal{X}}\frac{4}{m^2}\sum_{i=1}^m\mathbb{E}\Big[\Big|V(x) - \widehat{V}_i(x)\Big|^2\Big]\pi(dx), \\ &\quad \leq 4\int_{x\in\mathcal{X}}\frac{\mathbb{E}[\|v(W_1(0))\|_2^2]}{m}\pi(dx) \leq \frac{4\overline{\nu}^2}{m}, \quad (33)\end{aligned}$$

since $Var(\widehat{V}_i(x)) \leq \mathbb{E}[\|v(W_1(0))\|_2^2] \leq \overline{\nu}^2$ by Assumption 2 and $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}$ by Assumption 1. The extra factor is due to the symmetric initialization. By substituting (33) into (30), we have:

$$\begin{aligned}&\mathbb{E}\Big[\Delta_t\big(V(x_t) - \nabla_W^\top Q_0(x_t)\overline{W}\big); \mathcal{E}_t\Big] \\ &\qquad \leq \frac{4\overline{\nu}}{\sqrt{m}}\sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; \mathcal{E}_t]}.\end{aligned}$$

■

**Part (3)** Let

$$\overline{U}_i = a_i\frac{v(W_i(0))}{\sqrt{m}}, i \in [m],$$

with $\overline{U} = [\overline{U}_i]_{i\in[m]}$, which implies $\overline{W} = W(0) + \overline{U}$ [17]. Note that under symmetric initialization, $\nabla_W Q_0^\top(x)W(0) = Q_0(x) = 0$ for all $x \in \mathcal{X}$. Then, for any $t$, we have:

$$\begin{aligned}&\mathbb{E}[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top\overline{U}; \mathcal{E}_t] \\ &\qquad \leq \frac{4\overline{\nu}\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}}z_t, \quad (34)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top W(0); \mathcal{E}_t] \\ &\qquad \leq \frac{4\lambda\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}}z_t, \quad (35)\end{aligned}$$

with probability at least $1 - \delta$ over the random initialization.

*Proof:* In order to prove (34), we have the following bound by using Claim 5:

$$\begin{aligned}&\mathbb{E}_t[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top\overline{U}]\mathbb{I}_{\mathcal{E}_t} \\ &\quad \leq (1+\gamma)\|Q_t - V\|_\pi\|\nabla_W^\top Q_t\overline{U} - \nabla_W^\top Q_0\overline{U}\|_\pi \quad (36)\end{aligned}$$

For any $x \in \mathcal{X}$, we have:

$$\begin{aligned}&\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top\overline{U} \\ &= \sum_{i\in[m]}\Big(\mathbb{I}\{W_i^\top(0)x \geq 0\} - \mathbb{I}\{W_i^\top(t)x \geq 0\}\Big)\frac{v^\top(W_i(0))x}{m}.\end{aligned}$$

Let

$$S_x(t) = \Big\{i \in [m] : \mathbb{I}\{W_i^\top(0)x \geq 0\} \neq \mathbb{I}\{W_i^\top(t)x \geq 0\}\Big\}. \quad (37)$$

For any $x \in \mathcal{X}$ and $i \in S_x(t)$, we have:

$$|W_i^\top(0)x| \leq |W_i^\top(0)x - W_i^\top(t)x| \leq \|W_i(0) - W_i(t)\|_2,$$

since $i \in S_x(t)$ implies $W_i^\top(0)x$ and $W_i^\top(t)x$ have different signs. Therefore, we have the following relation:

$$\begin{aligned} S_x(t) &\subset \Big\{i \in [m] : |W_i^\top(0)x| \leq \|W_i(0) - W_i(t)\|_2\Big\}, \\ &\subset \Big\{i \in [m] : |W_i^\top(0)x| \leq \lambda/\sqrt{m}\Big\}, \end{aligned} \tag{38}$$

for any $t < \tau_1$. With this definition, we have:

$$\begin{aligned} &\big|\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top \overline{U}| \\ &\qquad \leq \frac{1}{m}\sum_{i\in[m]} \mathbb{I}\{i \in S_x(t)\}\bar{\nu} \leq \frac{4\bar{\nu}}{m}\widetilde{S}(x). \end{aligned} \tag{39}$$

since $v(w) \leq \bar{\nu}$ for any $w \in \mathbb{R}^d$ by Assumption 2, where

$$\widetilde{S}(x) = \sum_{i=1}^{m/2} \mathbb{I}\Big\{|W_i^\top(0)x| \leq \lambda/\sqrt{m}\Big\}, \tag{40}$$

for any $x \in \mathcal{X}$. By Claim 2, under $E_1 \cap \{t < \tau_1\}$, we have:

$$\frac{2\widetilde{S}(x)}{m} \leq \frac{\lambda}{\sqrt{m}} + \frac{\sqrt{2}\ell(m/2,\delta)}{\sqrt{m}}. \tag{41}$$

Therefore, we can bound (36) as follows:

$$\begin{aligned} &\mathbb{E}_t[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{U}]\mathbb{I}_{\mathcal{E}_t} \\ &\qquad \leq \frac{4\bar{\nu}(\lambda + \ell(m,\delta))}{\sqrt{m}}\|Q_t - V\|_\pi \mathbb{I}_{\mathcal{E}_t}. \end{aligned}$$

By taking expectation and using Cauchy-Schwarz inequality, we obtain:

$$\mathbb{E}[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{U}] \leq \frac{4\bar{\nu}\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}} z_t.$$

The above analysis builds on and improves the classical analysis of ReLU networks [13], [17] as it proposes uniform bounds over all $x \in \mathcal{X}$.

In order to prove (35), we use Claim 5 to obtain the following inequality:

$$\begin{aligned} &\mathbb{E}_t[\Delta_t\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top W(0)] \\ &\quad \leq 2\|Q_t - V\|_\pi \|\big(\nabla_W^\top Q_0 W(0) - \nabla_W^\top Q_t W(0)\|_\pi. \end{aligned} \tag{42}$$

For any $x \in \mathcal{X}$, we have:

$$\begin{aligned} &\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top W(0) \\ &= \frac{1}{\sqrt{m}}\sum_{i\in[m]} a_i\Big(\mathbb{I}\{W_i^\top(0)x \geq 0\} - \mathbb{I}\{W_i^\top(t)x \geq 0\}\Big) W_i^\top(0)x. \end{aligned}$$

Recall the definition of $S_x(t)$ in (37). By using triangle inequality:

$$\begin{aligned} &\big|\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top W(0)| \\ &\qquad \leq \frac{1}{\sqrt{m}}\sum_{i\in[m]} \mathbb{I}\{i \in S_x(t)\} \cdot |W_i^\top(0)x|. \end{aligned}$$

For any $x \in \mathcal{X}$ and $i \in S_x(t)$, we have:

$$|W_i^\top(0)x| \leq |W_i^\top(0)x - W_i^\top(t)x| \leq \|W_i(0) - W_i(t)\|_2,$$

since $i \in S_x(t)$ implies $W_i^\top(0)x$ and $W_i^\top(t)x$ have different signs. The correlation between $\mathbb{I}\{i \in S_x(t)\}$ and $\|W_i(t) - W_i(0)\|_2$ creates the main problem in the proof, which we resolve under the event $\{t < \tau_1\}$. For $t < \tau_1$, we have $\|W_i(0) - W_i(t)\|_2 \leq \lambda/\sqrt{m}$. Thus, we have:

$$\begin{aligned} \big|\big(\nabla_W Q_0(x) - \nabla_W Q_t(x)\big)^\top W(0)| &\leq \frac{\lambda}{m}\sum_{i\in[m]} \mathbb{I}\{i \in S_x(t)\}, \\ &\leq \frac{\lambda}{m}|S_x(t)| \leq \frac{4\lambda}{m}\widetilde{S}(x), \end{aligned}$$

where $\widetilde{S}(x)$ is defined in (40). Using Claim 2 similar to (41), under $E_1 \cap \{t < \tau_1\}$, we have:

$$\mathbb{E}[\Delta_t\big(\nabla_W Q_0(x_t) - \nabla_W Q_t(x_t)\big)^\top \overline{U}] \leq \frac{4\lambda\big(\lambda + \ell(m,\delta)\big)}{\sqrt{m}} z_t.$$

■

### B. Proximity of $\overline{Q}_T$ and $\frac{1}{T}\sum_{t<T} Q_t$

In this section, we will show that the output of Algorithm 1, $\overline{Q}_T(x) = Q(x; \frac{1}{T}\sum_{t<T} W(t), a)$, is close to $\frac{1}{T}\sum_{t<T} Q_t(x)$ in expectation, which will prove that $\overline{Q}_T$ achieves the target error. The idea is based on [10], and aims to use the linear approximation $\nabla_W^\top Q_0(x)\widehat{W}(T-1)$ as an auxiliary function to show the proximity of $\overline{Q}_T$ and $\frac{1}{T}\sum_{t<T} Q_t$.

*Proposition 3:* Let $\widetilde{W} \in \mathbb{R}^{md}$ be a (random) vector of parameters. Also, let $\widehat{Q}(x) = Q(x; \widetilde{W}, a)$ and $\widehat{Q}_0(x) = \nabla_W^\top Q_0(x)\widetilde{W}$ for any $x \in \mathcal{X}$, and the event $\mathcal{A} = \{\max_{i\in[m]} \|\widetilde{W}_i - W_i(0)\|_2 \leq \frac{\lambda}{\sqrt{m}}\} \cap E_1$. Then, we have the following inequality:

$$\mathbb{E}[\|\widehat{Q} - \widehat{Q}_0\|_\pi; \mathcal{A}] \leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}} \leq \frac{\epsilon}{2}.$$

Consequently, we have:

$$\mathbb{E}\Big[\Big\|\overline{Q}_T - \frac{1}{T}\sum_{t<T} Q_t\Big\|_\pi; \mathcal{E}_T\Big] \leq \epsilon. \tag{43}$$

*Proof:* First, note that the difference of $\widehat{Q}$ and $\widehat{Q}_0$ can be written as follows:

$$\begin{aligned} &|\widehat{Q}(x) - \widehat{Q}_0(x)| \\ &\leq \frac{1}{\sqrt{m}}\sum_{i\in[m]} \big|\mathbb{I}\{\widetilde{W}_i^\top x \geq 0\} - \mathbb{I}\{W_i^\top(0)x \geq 0\}\big| \cdot |\widetilde{W}_i^\top x|, \end{aligned}$$

for any $x \in \mathcal{X}$. Let

$$S_x = \Big\{i \in [m] : \mathbb{I}\{W_i^\top(0)x \geq 0\} \neq \mathbb{I}\{\widetilde{W}_i^\top x \geq 0\}\Big\}.$$

Then, we have:

$$\begin{aligned} |\widehat{Q}(x) - \widehat{Q}_0(x)| &\leq \frac{1}{\sqrt{m}}\sum_{i\in[m]} \mathbb{I}\{i \in S_x\}|\widetilde{W}_i^\top x| \\ &\leq \frac{1}{\sqrt{m}}\sum_{i\in[m]} \mathbb{I}\{i \in S_x\}\|\widetilde{W}_i - W_i(0)\|_2. \end{aligned}$$

since $i \in S_x$ implies $|\widetilde{W}_i^\top x| \leq |\widetilde{W}_i^\top x - W_i^\top(0)x| \leq \|\widetilde{W}_i - W_i(0)\|_2$. Similarly, we have: $|W_i^\top(0)x| \leq \|\widetilde{W}_i - W_i(0)\|_2$. Then, we have:

$$\begin{aligned}|\widehat{Q}(x) - \widehat{Q}_0(x)|\mathbb{I}_{\mathcal{A}} &\leq \frac{\lambda}{m}|S_x|\mathbb{I}_{\mathcal{A}} \\ &\leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}}.\end{aligned}$$

Taking the expectation and using Jensen's inequality, we have:

$$\begin{aligned}\mathbb{E}[\|\widehat{Q} - \widehat{Q}_0\|_\pi; \mathcal{A}] &\leq \sqrt{\mathbb{E}[\|\widehat{Q}_T - \widehat{Q}_0\|_\pi^2; \mathcal{A}]} \\ &\leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}},\end{aligned}$$

which concludes the proof of the first claim.

In order to prove the second claim, consider $\widetilde{W} = \widehat{W}(T-1) = \frac{1}{T}\sum_{t<T} W(t)$, and note that $\widehat{W}(T-1) \in \mathcal{F}_{T-1}$ and $\mathcal{E}_T \subset \mathcal{A}$ by definition. Therefore, the first part implies the following:

$$\mathbb{E}[\|\overline{Q}_T - \nabla_W^\top Q_0 \widehat{W}(T-1)\|_\pi; \mathcal{E}_T] \leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}}. \quad (44)$$

with the usual notation $\nabla_W^\top Q_0 \widehat{W}(T-1) = \big[\nabla_W^\top Q_0(x)\widehat{W}(T-1)\big]_{x\in\mathcal{X}}$. Finally, we have:

$$\begin{aligned}\mathbb{E}[\|\frac{1}{T}\sum_{t<T} Q_t - \nabla_W^\top Q_0 \widehat{W}(T-1)\|_\pi; \mathcal{E}_T] \\ \leq \frac{1}{T}\sum_{t<T}\mathbb{E}[\|Q_t - \nabla_W^\top Q_0 W(t)\|_\pi; \mathcal{E}_T],\end{aligned}$$

by Jensen's inequality. For any $t < T$, letting $\widetilde{W} = W(t)$, and noting that $\mathcal{E}_T \subset \mathcal{A}$, we have $\mathbb{E}[\|Q_t - \nabla_W^\top Q_0 W(t)\|_\pi; \mathcal{E}_t] \leq \frac{\lambda(\lambda+\ell(m,\delta))}{\sqrt{m}}$ by using the first part of the proposition, which implies:

$$\mathbb{E}[\|\frac{1}{T}\sum_{t<T} Q_t - \nabla_W^\top Q_0 \widehat{W}(T-1)\|_\pi; \mathcal{E}_T] \leq \frac{\lambda(\lambda + \ell(m,\delta))}{\sqrt{m}}. \quad (45)$$

Using (44), (45) and triangle inequality together, we conclude that

$$\mathbb{E}[\|\frac{1}{T}\sum_{t<T} Q_t - \overline{Q}_T\|_\pi; \mathcal{E}_T] \leq \epsilon,$$

with the choice of parameters in Theorem 1. ∎

## Appendix II
## Proof of Theorem 2

The proof of Theorem 2 consists of the same steps as Theorem 1, but it is simpler because the growth of $\|W(t) - \overline{W}\|_2$ is controlled by the max-norm constraint. In the first step, we will prove a Lyapunov drift bound.

### A. Lyapunov Drift Bound

First, note that for any $R > 0$ and $m \in \mathbb{N}$,

$$\mathcal{G}_{m,R} = \Big\{w \in \mathbb{R}^{md} : \|W_i(0) - w_i\|_2 \leq \frac{R}{\sqrt{m}}, \forall i \in [m]\Big\},$$

is the Cartesian product of convex sets $\mathcal{G}_{m,R}^i$, which is convex. This leads to the following result.

*Lemma 3:* For any $t \geq 0$ and $R \geq \bar{\nu}$, we have the following inequalities:

$$\begin{aligned}\mathbb{E}[\|W(t+1) - \overline{W}\|_2^2; E_1] &\leq \mathbb{E}[\|W(t) - \overline{W}\|_2^2; E_1] \\ &\quad - 2\alpha(1-\gamma)z_t^2 + \alpha^2(1+2R)^2 \\ &\quad + 8\alpha z_t\Big(\frac{\bar{\nu} + \big(\bar{\nu} + R\big)\big(R + \ell(m,\delta)\big)}{\sqrt{m}}\Big),\end{aligned} \quad (46)$$

where $\overline{W}$ is as defined in (12), $z_t = \sqrt{\mathbb{E}[\|Q_t - V\|_\pi^2; E_1]}$.

*Proof:* First, note that $W(t+1) = \Pi_{\mathcal{G}_{m,R}} W(t+1/2)$ by the update rule in (9), and $\mathcal{G}_{m,R}$ is a convex set. Also, note that $R \geq \bar{\nu}$ implies $\overline{W} \in \mathcal{G}_{m,R}$. Therefore, we have:

$$\begin{aligned}\|W(t+1) - \overline{W}\|_2^2 &= \|\Pi_{\mathcal{G}_{m,R}} W(t+1/2) - \Pi_{\mathcal{G}_{m,R}}\overline{W}\|_2^2, \\ &\leq \|W(t+1/2) - \overline{W}\|_2^2,\end{aligned}$$

which follows since projection is a non-expansive operation for convex subsets. Since $W(t+1/2) = W(t) + \alpha g_t$ and $\|g_t\|_2 \leq 1 + 2R$ by (19), we have:

$$\begin{aligned}\mathbb{E}_t\|W(t+1) - \overline{W}\|_2^2 \leq \|W(t) - \overline{W}\|_2^2 + 2\alpha\mathbb{E}_t[g_t^\top](W(t) - \overline{W}) \\ + \alpha^2(1+2R)^2.\end{aligned}$$

Then, the proof follows by multiplying both sides by $\mathbb{I}_{E_1}$, taking expectation, and using Proposition 1 with $\lambda$ replaced by $R$ since $\|W_i(t) - W_i(0)\|_2 \leq R/\sqrt{m}$ for all $i \in [m]$ and $\tau_1 = \infty$. ∎

### B. Error Bound

Note that by the choices of step-size $\alpha$ and network width $m$, we have:

$$\alpha^2(1+2R)^2 = \alpha(1-\gamma)\epsilon^2,$$

and

$$\frac{\bar{\nu} + \big(\bar{\nu} + R\big)\big(R + \ell(m,\delta)\big)}{\sqrt{m}} \leq \epsilon(1-\gamma)/4.$$

Using these in Lemma 3, we have:

$$\begin{aligned}\mathbb{E}[\|W(t+1) - \overline{W}\|_2^2; E_1] &\leq \mathbb{E}[\|W(t) - \overline{W}\|_2^2; E_1] \\ &\quad - \alpha(1-\gamma)\Big(z_t - \epsilon\Big)^2 + 2\alpha(1-\gamma)\epsilon^2.\end{aligned}$$

By telescoping sum over $t = 0, 1, \ldots, T-1$, the above inequality yields:

$$\begin{aligned}\frac{1}{T}\sum_{t<T}(z_t - \epsilon)^2 &\leq \frac{\mathbb{E}[\|W(0) - \overline{W}\|_2^2; E_1]}{\alpha(1-\gamma)T} + 2\epsilon^2, \\ &\leq \frac{\bar{\nu}^2}{\alpha(1-\gamma)T} + 2\epsilon^2.\end{aligned}$$

By using Jensen's inequality,

$$\Big(\frac{1}{T}\sum_{t<T} z_t - \epsilon\Big)^2 \leq \frac{\bar{\nu}^2}{\alpha(1-\gamma)T} + 2\epsilon^2.$$

The above inequality yields:

$$\begin{aligned}\frac{1}{T}\sum_{t<T}\mathbb{E}[\|Q_t - V\|_\pi; E_1] &\leq \frac{1}{T}\sum_{t<T} z_t \\ &\leq \frac{\bar{\nu}}{\sqrt{\alpha(1-\gamma)T}} + 3\epsilon.\end{aligned}$$

We conclude the proof by using Proposition 3.

## APPENDIX III
## ANALYSIS OF MN-NTD UNDER MARKOVIAN SAMPLING

In Theorem 2, we analyzed the performance of MN-NTD under iid sampling, where $s_t \sim \pi$ independently for any $t$, and $s_t' \sim P(s_t, \cdot)$. The convergence of TD learning under Markovian sampling is established by using mixing time analysis [7], [38], [47]. In the following, we provide a finite-time analysis of MN-NTD under Markovian sampling by directly adapting the analysis in [7], [47].

We make the following uniform mixing assumption, which is commonly used in the analysis of Markovian sampling [7], [47].

*Assumption 3:* There exist constants $\varsigma > 0$ and $\rho \in (0,1)$ such that

$$\sup_{s\in\mathcal{S}} \|\mathbb{P}(s_t \in \cdot | s_0 = s) - \pi(\cdot)\|_1 \leq \varsigma \rho^t, \tag{47}$$

for any $t \geq 0$.

Under Assumption 3, there is a mixing time $\tau_{mix}$ such that $\tau_{mix} = \min\{t \in 0 : \varsigma\rho^t \leq \frac{\epsilon^2(1-\gamma)}{8R(1+2R)}\}$ [7].

The performance of MN-NTD under Markovian sampling with $s_0 \sim \pi$ is as follows.

*Theorem 3:* Under Assumptions 1-3, for any $\epsilon > 0$, $\delta \in (0,1)$ and $R > \bar{\nu}$, MN-NTD with width

$$m > \frac{64\Big(\bar{\nu} + \big(R + \ell(m,\delta)\big)\big(\bar{\nu} + R + 4R(1+2R)\big)\Big)^2}{(1-\gamma)^2\epsilon^2},$$

and step-size

$$\alpha = \frac{\epsilon^2(1-\gamma)}{c(R,\tau_{mix})},$$

satisfies the following under Markovian sampling:

$$\mathbb{E}\Big[\|\overline{Q}_T - V\|_\pi ; E_1\Big] \leq \frac{\sqrt{c(R,\tau_{mix})}\bar{\nu}}{\epsilon(1-\gamma)\sqrt{T}} + 6\epsilon,$$

where

$$c(R,\tau_{mix}) = 4(1+2R)^2 + 8(1+2R)(1+10R)\tau_{mix},$$

and $E_1 \in \mathcal{F}_{init}$ holds with probability at least $1-\delta$.

*Proof:* The proof builds on the analysis in [47]. For $y = (x, x') \in \mathcal{X}^2$ and $W \in \mathbb{R}^{m\times d}$, let

$$g(W,y) = \Big(r(x) + \gamma Q(x';W,a) - Q(x;W,a)\Big)\nabla Q(x;W,a),$$

and

$$h(W,y) = \Big(r(x) + \gamma Q(x';W,a) - Q(x;W,a)\Big)\nabla Q_0(x).$$

Also, define

$$Z(W,y) = \big[h(W,y) - \overline{h}(W)\big]^\top (W - \overline{W}),$$

where $\overline{h}(W) = \mathbb{E}_{y\sim\mu\otimes P}[h(W,y)]$. Let $h_t(W) = h(W,y_t)$ and $Z_t(W) = Z(W,y_t)$ where $y_t = (x_t, x_{t+1})$. By following identical steps as Lemma 3, we obtain the following drift bound under Markovian sampling:

$$\begin{aligned}\|W(t+1) - \overline{W}\|_2^2 \leq{}& \|W(t) - \overline{W}\|_2^2 + \alpha^2(1+2R)^2 \\ &+ 2\alpha\overline{g}^\top(W(t))(W(t) - \overline{W}) \\ &+ 2\alpha\big(g_t(W(t)) - h_t(W(t))\big)^\top(W(t) - \overline{W}) \quad (48)\\ &+ 2\alpha(h_t(W(t)) - \overline{h}(W(t)))^\top\big(W(t) - \overline{W}\big) \quad (49)\\ &+ 2\alpha(\overline{h}(W(t)) - \overline{g}(W(t)))^\top(W(t) - \overline{W}). \quad (50)\end{aligned}$$

The terms (48)-(50) are due to Markovian sampling, and the rest of the drift bound is identical to the iid sampling case considered in Theorem 2.

First, by Lemma 2 in [11], we have:

$$\begin{aligned}&\Big(g_t(W(t)) - h_t(W(t))\Big)^\top(W(t) - \overline{W}) \\ &\qquad\leq \frac{2R}{\sqrt{m}}(1+2R)(R + \ell_0(m,\delta)), \quad (51)\end{aligned}$$

and

$$\begin{aligned}&\Big(\overline{h}(W(t)) - \overline{g}(W(t))\Big)^\top(W(t) - \overline{W}) \\ &\qquad\leq \frac{2R}{\sqrt{m}}(1+2R)(R + \ell_0(m,\delta)), \quad (52)\end{aligned}$$

simultaneously under the event $E_1$.

Note that the term in (49) is $2\alpha Z_t(W(t))$. In the following, we will bound $\mathbb{E}[Z_t(W(t)); E_1]$ by using Lemmas 8 and 9 in [7].

Let $W, W' \in \mathbb{R}^{m\times d}$ be such that

$$\max_{i\in[m]} \max\{\|W_i - W_i(0)\|_2, \|W_i' - W_i(0)\|_2\} \leq \frac{R}{\sqrt{m}}. \tag{53}$$

Then, by Lemma 2 in [47], we have

$$\begin{aligned}|Z_t(W) - Z_t(W')| \leq{}& \|h_t(W) - h_t(W')\|_2 \cdot \|W - \overline{W}\|_2 \\ &+ \|\overline{h}(W) - \overline{h}(W')\|_2 \cdot \|W - \overline{W}\|_2 \\ &+ \|h_t(W) - h_t(W')\|_2 \cdot \|W - W'\|_2.\end{aligned}$$

Since $Q(x;W,a)$ is 1-Lipschitz continuous in $W$ for any $x \in \mathbb{R}^d$ with $\|x\|_2 \leq 1$, we have $\|h_t(W) - h_t(W')\|_2 \leq 2\|W - W'\|_2$ and $\|\overline{h}(W) - \overline{h}(W')\|_2 \leq 2\|W - W'\|_2$. Also, since $\max\{\|h_t(W)\|_2, \overline{h}(W)\|_2\} \leq 1 + 2R$, we obtain the following inequality:

$$|Z_t(W) - Z_t(W')| \leq \Big(4(R+\bar{\nu}) + 1 + 2R\Big)\|W - W'\|_2. \tag{54}$$

Then, as a consequence of max-norm projection, the iterates $\{W(t) : t \in [T]\}$ satisfy (53), thus (54) is also satisfied. This implies the following:

$$\begin{aligned}Z_t(W(t+1)) - Z_t(W(t)) \leq{}& 2\Big(4(R+\bar{\nu}) + 1 + 2R\Big) \\ &\times \|W(t+1) - W(t)\|_2,\end{aligned}$$

which implies that

$$Z(W(t+1)) \leq Z(W(t)) + 2\alpha(1+10R)(1+2R),$$

since $R > \bar{\nu}$ and $\|W(t+1) - W(t)\|_2 \leq \alpha(1+2R)$. Then, for any $\tau \geq 0$, we have:

$$\mathbb{E}_0 Z_t(W(t)) \leq \mathbb{E}_0 Z_t(W(t-\tau)) + 2\alpha(1+10R)(1+2R)\tau, \tag{55}$$

where the expectation is taken over the samples conditioned on the random initialization. Recall that $\tau_{mix} = \inf\{t \geq 1 : \varsigma\rho^t \leq \frac{\epsilon^2(1-\gamma)}{8R(1+2R)}\}$ is the mixing time of the underlying Markov chain.

- If $t \leq \tau_{mix}$, letting $\tau = t$, we have:

$$\begin{aligned}\mathbb{E}_0[Z_t(W(t))] &\leq \mathbb{E}_0[Z_t(W(0))] + \alpha(1+2R)(1+10R)\tau_{mix},\\ &= \alpha(1+2R)(1+10R)\tau_{mix}, \qquad (56)\end{aligned}$$

since $(a, W(0))$ is independent of the underlying Markov chain $\{x_t : t \geq 0\}$, thus $\mathbb{E}_0[Z_t(W(0))] = 0$ for any $t$ by the definition of $Z_t$.

- If $t > \tau_{mix}$, we let $\tau = \tau_{mix}$ in (55). Then, we have:

$$\mathbb{E}_0 Z_t(W(t)) \leq \alpha(1+2R)(1+10R)\tau_{mix} + \mathbb{E}_0 Z_t(W(t-\tau_{mix})).$$

In order to bound $\mathbb{E}_0 Z_t(W(t-\tau_{mix}))$, we use Lemma 9 in [7], which implies that

$$\begin{aligned}&|\mathbb{E}_0 Z(W(t-\tau_{mix}), y_t) - \mathbb{E}_0 Z(W'(t-\tau_{mix}), y_t')|\\ &\qquad \leq 2\sup_{W,y}|Z(W,y)|\frac{\epsilon^2(1-\gamma)}{8R(1+2R)},\end{aligned}$$

for statistically independent $W'(t-\tau_{mix}) \overset{d}{=} W(t-\tau_{mix})$ and $y_t' \overset{d}{=} y_t = (x_t, x_{t+1})$, where $\overset{d}{=}$ denotes equality in distribution. For such $W'(t-\tau_{mix})$ and $y_t'$, we have $\mathbb{E}_0[Z(W'(t-\tau_{mix}), y_t')|Z(W'(t-\tau_{mix})] = 0$, which implies that

$$\mathbb{E}_0 Z_t(W(t-\tau_{mix})) \leq \frac{\epsilon^2(1-\gamma)}{4},$$

and

$$\mathbb{E}_0 Z_t(W(t)) \leq \alpha(1+2R)(1+10R)\tau_{mix} + \frac{\epsilon^2(1-\gamma)}{4}. \quad (57)$$

Then, from (55) and (57), for any $t \geq 0$, we have:

$$\mathbb{E}_0 Z_t(W(t)) \leq \alpha(1+2R)(1+10R)\tau_{mix} + \frac{\epsilon^2(1-\gamma)}{4},$$

under the event $E_1$. Now that we have established upper bounds for (48)-(50), we have the following drift inequality:

$$\begin{aligned}\mathbb{E}_0\|W(t+1) - \overline{W}\|_2^2 &\leq \mathbb{E}_0\|W(t) - \overline{W}\|_2^2\\ &+ 2\alpha\mathbb{E}_0[\overline{g}^\top(W(t))(W(t) - \overline{W})]\\ &+ 8\alpha\frac{R}{\sqrt{m}}(1+2R)(R + \ell_0(m,\delta))\\ &+ 2\alpha^2(1+2R)(1+10R)\tau_{mix} + \frac{\alpha\epsilon^2(1-\gamma)}{2}\\ &+ \alpha^2(1+2R)^2,\end{aligned}$$

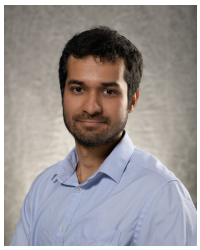
under the event $E_1 \in \mathcal{F}_{init}$. Note that the step-size choice ensures the following:

$$\alpha^2\Big((1+2R)^2 + 2(1+2R)(1+10R)\tau_{mix})\Big) \leq \frac{\alpha\epsilon^2(1-\gamma)}{2}.$$

The proof follows by taking expectation under the event $E_1$, using the bound on $\mathbb{E}[\overline{g}^\top(W(t))(W(t) - \overline{W}); E_1]$ by using Proposition 1, telescoping sum over $t = 0, 1, \ldots, T-1$, and using Proposition 3, similar to the proofs of Theorems 1 and 2.

■

**Semih Cayci** (S '12) received his PhD degree in electrical and computer engineering from the Ohio State University, Columbus, OH in 2020. From 2020-2022, he was a Postdoctoral Research Associate at the University of Illinois at Urbana-Champaign. Currently, he is a Juniorprofessor (W1 tenure-track W2) for Mathematics of Machine Learning in the Department of Mathematics at RWTH Aachen University. His research interests include learning theory, optimization and applied probability.



**Siddhartha Satpathi** received his B.Tech. and M.Tech. in electronics and electrical communication engineering from Indian Institute of Technology, Kharagpur in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Illinois at Urbana-Champaign. His research interests include optimization and machine learning.

**Niao He** (S'10-M'15) received B.S. in Mathematics from University of Science and Technology of China in 2010 and obtained M.S. in Computational Science and Engineering and Ph.D. in Operations Research both from Georgia Institute of Technology, USA in 2015.

From 2016-2020, she was an assistant professor with the Coordinated Science Laboratory and Department of Industrial Systems and Engineering at University of Illinois at Urbana-Champaign, IL, USA. Currently, she is an assistant professor in the Department of Computer Science at ETH Zurich, Switzerland. Her research interests include optimization, reinforcement learning, and machine learning.

**R. Srikant** (S '90-M '91-SM '01-F '06) received his B.Tech. from the Indian Institute of Technology, Madras in 1985, his M.S. and Ph.D. from the University of Illinois in 1988 and 1991, respectively, all in Electrical Engineering. He was a Member of Technical Staff at AT&T Bell Laboratories from 1991 to 1995. He is currently with the University of Illinois at Urbana-Champaign, where he is the co-Director of the C3.ai Digital Transformation Institute, a Grainger Distinguished Chair in Engineering and a Professor in the Department of Electrical and Computer Engineering and the Coordinated Science Lab.

He is the recipient of the 2015 INFOCOM Achievement Award, the 2019 IEEE Koji Kobayashi Computers and Communications Award and the 2021 ACM SIGMETRICS Achievement Award. He has also received several Best Paper awards including the 2015 INFOCOM Best Paper Award, the 2017 Applied Probability Society Best Publication Award, and the 2017 WiOpt Best Paper award. He was the Editor-in-Chief of the IEEE/ACM Transactions on Networking from 2013-2017. His research interests include applied probability, machine learning, stochastic control and communication networks.